

KATHOLIEKE UNIVERSITEIT
LEUVEN

Arenberg Doctoral School of Science, Engineering & Technology
Faculty of Engineering
Department of Electrical Engineering (ESAT)

Design of Clinical Decision Support Systems for Cancer based upon Clinical and Molecular Data

Anneleen Daemen

Dissertation presented in partial
fulfillment of the requirements for
the degree of Doctor
in Engineering

May 2010

Design of Clinical Decision Support Systems for Cancer based upon Clinical and Molecular Data

Anneleen Daemen

Jury:
Prof. dr. ir. P. Van Houtte, chairman
Prof. dr. ir. B. De Moor, promotor
Prof. dr. dr. D. Timmerman, promotor
Prof. dr. ir. J.A.K. Suykens
Prof. dr. E. Waelkens
Prof. dr. ir. T. De Bie (Univ. of Bristol)
Prof. dr. L. Dehaspe

Dissertation presented in partial fulfillment of the requirements for the degree of Doctor in Engineering

May 2010

© Katholieke Universiteit Leuven – Faculty of Engineering
Address, B-3001 Leuven (Belgium)

Alle rechten voorbehouden. Niets uit deze uitgave mag worden vermenigvuldigd en/of openbaar gemaakt worden door middel van druk, fotocopie, microfilm, elektronisch of op welke andere wijze ook zonder voorafgaande schriftelijke toestemming van de uitgever.

All rights reserved. No part of the publication may be reproduced in any form by print, photoprint, microfilm or any other means without written permission from the publisher.

Legal depot number D/2010/7515/50
ISBN number 978-94-6018-211-2

Voorwoord

Tijdens de blok in het najaar van 2005 besefte ik plots dat als ik met een doctoraat wilde aanvangen, ik niet bij de pakken mocht blijven zitten. Ik kwam immers in aanmerking voor een FWO beurs waarvan de deadline binnen de maand was. Na een mailtje naar Prof. Bart De Moor kon ik tussen Kerstmis en Nieuwjaar bij hem terecht. Vanaf toen is alles heel snel gegaan. Het onderwerp lag nog niet vast, maar ik wist wel dat de richting bioinformatica me zou liggen door de mix van wiskunde en geneeskunde, twee onderwerpen die me altijd al hadden geboeid. Mijn eerste dank gaat dan ook uit naar mijn promoter. Bart, bedankt voor je vertrouwen, de mogelijkheid om te groeien, met vallen en opstaan mijn eigen weg te gaan, en mijn licht te kunnen opsteken op tal van conferenties.

Vervolgens wil ik ook mijn oprechte dank betuigen aan mijn co-promoter Prof. Dirk Timmerman. Bedankt Dirk, voor je onontbeerlijke steun, fijne samenwerking en de talrijke interessante projecten waar ik het genoegen had bij betrokken te worden. De leden van mijn begeleidingscommissie, Prof. Johan Suykens en Prof. Etienne Waelkens, ben ik dank verschuldigd voor hun steun gedurende mijn doctoraat en het doornemen van dit proefschrift. Johan zou ik in het bijzonder willen bedanken voor zijn cruciale bijdragen op methodologisch vlak. Prof. Paul Van Houtte, als voorzitter, Prof. Luc Dehaspe en Prof. Tijl De Bie zou ik willen bedanken voor hun bereidheid te zetelen in mijn examencommissie. Het fonds voor wetenschappelijk onderzoek (FWO) ben ik tenslotte dankbaar voor de financiële steun gedurende de voorbije vier jaar.

Aangezien mijn doctoraat gekenmerkt werd door veel gevarieerde en interessante samenwerkingen, wil ik naast mijn jury iedereen bedanken die op professioneel vlak tot dit doctoraat heeft bijgedragen. In het bijzonder wil ik de volgende personen bedanken: Prof. Karin Haustermans, Prof. Jean-Pascal Machiels en Annelies Debucquoy voor de uitdagende samenwerking rond multimodale data van darmkanker; Ann Smeets voor de fijne samenwerking rond borstkanker; Prof. Jos van Pelt, Prof. Chris Verslype, Louis Libbrecht en Hannah Van Malenstein voor het mooie project rond leverkanker; Pim French, Andrew Stubbs, Lonneke Gravendeel, Johan de Rooi en Prof. Peter van der Spek voor de samenwerking

rond hersentumoren (het was steeds weer een plezier om naar Rotterdam te komen); Karin Leunen en Prof. Eric Legius voor de interessante projecten rond array CGH data; Prof. Ignace Vergote, Prof. Etienne Waelkens, Isabelle Cadron en Toon van Gorp voor de fijne samenwerking omtrent proteomics data; Ben Van Calster, Vanya Van Belle en Prof. Sabine Van Huffel voor de statistische noot; Prof. Dirk Timmerman, Prof. Lil Valentin, Prof. Tom Bourne, Caroline Van Holsbeke en het hele IOTA team om deel te mogen uitmaken van deze multicentrische studie. I would also like to thank Cecilia Bottomley, Emma Kirk, Anne Pexsters, Thierry Van den Bosch, Shabana Bora, Aris Papageorghiou, Prof. Tom Bourne and Prof. Christoph Lees for the nice collaborations on pregnancy and growth, and especially for the diners, drinks and chats at conferences and after many meetings.

Vervolgens wil ik Prof. Bart De Moor, Prof. Dirk Timmerman en Prof. Chris Verslype bedanken voor hun steun en bijdrage tot het behalen van een BAEF (Belgian American Educational Foundation) beurs. Mede dankzij jullie kan ik nu mijn licht gaan opsteken aan een Amerikaanse universiteit.

Een doctoraat zou niet tot een goed einde gebracht kunnen worden zonder een goede sfeer in de toren. Raf, Thomas, Karen, Tim, Wout, Pieter, Tom, Marcelo, Riet, Liesbeth, Lieven, Mauricio, Tunde, Peter K, Ernesto, Jiqiu en alle anderen, bedankt voor de momenten van ontspanning, de vele lunchpauze's (al ontbrak het mij naar het einde toe wat vaker aan tijd), de sista-weekends en de nachtelijke avonturen op gezamenlijke conferenties. In het bijzonder zou ik Olivier willen bedanken voor de begeleiding, vele discussies en vruchtbare resultaten. Bedankt dat je zo sterk focuste op output en mij na enkele maanden reeds aanzette hetzelfde te doen. Wie weet wanneer ik anders pas de moed had gevonden mijn eerste paper te submitten. I would also like to thank Marco, Fabian and Shi for the, somehow, cross-disciplinary conversations and collaborations that helped me to diverge not completely towards clinical applications. Tenslotte ook nog een woordje van dank aan Ida, Ilse, John, Mimi, Lut, Maarten en Liesbeth, die altijd klaarstonden om de praktische kant van een doctoraat in goede banen te leiden.

Naast het werk hebben ook familie en vrienden - al dan niet bewust - hun inbreng tot dit doctoraat gehad. Enkele vrienden zou ik in het bijzonder willen bedanken voor de amusante verjaardagsfeestjes en weekendjes: Sofie, Guy, Jef, Wouter, Pieter, Nancy en Davy.

Graag zou ik ook Marietje Brouns willen bedanken voor al haar moeite om mijn tekst grondig na te lezen.

Mijn ouders ben ik eeuwige dank verschuldigd voor al de kansen die ze mij gegeven hebben, en vooral om mij steeds datgene te laten doen dat ik het belangrijkste vond. Mama, papa, broer Bart, Nancy en broertje Dirk, bedankt voor alles!

En tenslotte Raf, jij was er bij vanaf het begin van mijn doctoraat en hebt me al die tijd op wonderbaarlijke wijze gesteund. Al snel bleek ook dat we gelijkaardige doelen nastreven zodat er nu een aantal (hopelijk) onvergetelijke jaren in Amerika op ons liggen te wachten.

Anneleen

Leuven
Mei 2010

Abstract

The tsunami of data with clinical parameters and information from the genome, transcriptome, proteome and metabolome requires advanced mathematical approaches to transform the huge amount of data in meaningful support for clinicians. This dissertation is therefore directed towards analytical tools for the heterogeneous disease cancer, in order to move from bench to bedside with decision support systems in daily clinical practice.

Because none of the available data types provides a comprehensive understanding of the complex biology of cancer, complementary levels of biological information need to be integrated in order to efficiently translate cancer genome information into molecular markers and therapeutics. The way we approach data fusion differs from what recently has been described in the literature. We present a kernel-based integration framework in which multiple layers of experimental data are combined in one mathematical model for the development of more homogeneous classifiers.

An extra layer of processing is required for data types that have not been used for classification so far or that would benefit from an improved modeling to fully exploit its content. For array CGH data, we present a method for the conversion of copy number variations into features usable for classification. As histopathology will remain a cornerstone in cancer decision making, we improve modeling of clinicopathological parameters with a kernel function specifically adapted to the characteristics of clinical data.

Finally, our data integration approach can be improved with non-patient related data obtained from databases on different aspects of biological systems. We describe how prior biological knowledge can be incorporated in classification models, by exploiting relations between genes with similar functions but active in alternative pathways based on spectral graph theory. Our data integration approach, the extension towards more complex data types and the incorporation of prior interactome knowledge contribute to improved clinical decision support and deciphering biological deregulated processes in which key markers are involved.

Korte Inhoud

Geavanceerde wiskundige methoden zijn onontbeerlijk om de enorme hoeveelheid aan klinische data en informatie afkomstig uit het centrale dogma van de moleculaire biologie om te vormen tot bruikbare en nuttige gegevens voor geneeskundigen. Dit proefschrift is dan ook gericht op het introduceren van klinische besluitvormingssystemen in de dagelijkse behandeling van kanker.

Doordat de beschikbare patiëntgegevens niet volstaan om de complexe biologie van kanker te ontrafelen, kunnen moleculaire merkers slechts geïdentificeerd worden door het integreren van biologisch complementaire data. De manier waarop wij het probleem van gegevensintegratie hebben benaderd, verschilt van wat recent in de literatuur is verschenen. Wij stellen namelijk een kernel-gebaseerde methode voor om meerdere types van patiënt-gerelateerde, experimentele data in een homogeen model te combineren.

Voor nieuwere data-types die nog niet voor classificatie zijn aangewend, is een extra verwerking vereist vooraleer onze integratie-methode toepasbaar is. Daarom introduceren we voor genomische data een methode waarmee variaties in het aantal genkopieën bruikbaar worden voor classificatie. Aangezien histopathologie belangrijk zal blijven voor de besluitvorming van kanker, verbeteren we eveneens de modellering van klinische gegevens door de lineaire kernelfunctie aan te passen aan de specifieke kenmerken van klinische data. Op deze manier kan er meer informatie geëxtraheerd worden vooraleer ze te combineren met andere gegevens.

Tenslotte kan classificatie verbeterd worden aan de hand van niet-patiëntgerelateerde, biologische kennis via spectrale grafentheorie. Hiervoor exploreren we relaties tussen genen met gelijkaardige functies maar die actief zijn in alternatieve pathways. Als besluit kunnen we stellen dat onze methode voor data-integratie, de uitbreiding naar complexere data-types en het incorporeren van interactoom-gegevens, bijdraagt tot een verbeterde klinische besluitvorming en het ontrafelen van biologisch gedereguleerde processen.

Acronyms

A	Adenine
AUB	Abnormal uterine bleeding
AUC	Area under the ROC curve
BER	Balanced error rate
BRCA	Breast cancer early onset gene
C	Cytosine
CA 125	Cancer Antigen 125
cDNA	Complementary DNA
CDS	Clinical decision support
CGH	Comparative genomics hybridization
CNV	Copy number variation
CRL	Crown-rump length
CRM	Circumferential margin involvement
CT	Computed tomography
CV	Cross-validation
DEDS	Differential expression via distance synthesis
DNA	Deoxyribonucleic acid
DDI	Domain-domain interaction
EGFR	Epidermal growth factor receptor
EHMN	Edinburgh human metabolic network
ELISA	Enzyme-linked immunosorbent assay
EP	Ectopic pregnancy
ER	Estrogen receptor
FDA	U.S. Food and drug administration
FLDA	Functional linear discriminant analysis
FN	False negatives
FP	False positives
FS	Feature selection
G	Guanine
GO	Gene Ontology
GOTM	Gene Ontology Tree Machine

GSEA	Gene set enrichment analysis
hCG	human chorionic gonadotropin
HER2	Human epidermal growth factor receptor 2
HI-DID-IT	HIgh-DImensional Data Integration Toolbox
HMM	hidden Markov Model
HUGO	Human Genome Organisation
IOTA	International ovarian tumor analysis
IUP	intrauterine pregnancy
J	Youden's index
KEGG	Kyoto encyclopedia of genes and genomes
LC-ESI	liquid chromatography coupled to electrospray ionization
LDA	Linear discriminant analysis
LMD	Laser microdissection
LN	Lymph node
LOO	Leave-one-out
LR	Logistic regression
LR−	Negative likelihood ratio
LR+	Positive likelihood ratio
LS-SVM	Least Squares Support Vector Machine
MALDI	Matrix-assisted laser desorption and ionization
MAS5.0	Microarray Analysis Suite 5.0 software (Affymetrix)
MetaCyc	Multiorganism database of metabolic pathways and enzymes
MI	Mutual information
MINDACT	Microarray for node negative disease may avoid chemotherapy
miRNA	microRNA
MKL	Multiple kernel learning
ML	Machine learning
MRI	Magnetic resonance imaging
mRNA	messenger RNA
MS	Mass spectrometry
MSE	Mean squared error
MSigDB	Molecular Signatures Database
NB	Naïve bayes
NET	Network-constrained regularization
NF	Number of features
NIH	National Institutes of Health
NSCLC	Non-small cell lung carcinoma
OPHID	Online predicted human interaction database
OR	Odds ratio
PAPP-A	Pregnancy associated plasma protein A
PGDB	Pathway genome database
PPI	Protein-protein interaction

PR	Progesterone receptor
PSA	Prostate-specific antigen
psd	Positive semidefinite
PUL	Pregnancy of unknown location
QP	Quadratic programming
QqTOF	Quadrupole time-of-flight
RBF	Radial basis function
RMA	Robust multichip analysis
RNA	Ribonucleic acid
ROC	Receiver operating characteristics
SAM	Significance analysis for microarrays
SELDI	Surface-enhanced laser desorption and ionization
SNP	Single nucleotide polymorphism
STRING	Search tool for recurring instances of neighboring genes
SVM	Support Vector Machine
T	Thymine
TAILORx	Trial assigning individualized options for treatment [Rx]
TF	Transcription factor
TN	True negatives
TOF	Time-of-flight
TP	True positives
TP53	Tumor protein 53
qRT-PCR	quantitative reverse transcriptase polymerase chain reaction
U	Uracil
UniDomInt	Unified domain interaction database
US	Ultrasound
wLS-SVM	weighted Least Squares Support Vector Machine

Symbols

x^i	pattern, sample, instance or observation
y_i	label, target or outcome
i	index parameter for the training data, $i = 1, \ldots, N$
N	size of the training data
N_N	size of the training data in the negative class
N_P	size of the training data in the positive class
k, l	index parameter for the number of variables, $k, l = 1, \ldots, p$
p	number of variables
X	$N \times p$ matrix
z	variable
f	real-valued function
Φ	feature map
k	kernel function
τ	intercept constant of the polynomial kernel function
d	degree of the polynomial kernel function
σ	width of the RBF kernel function
K	kernel or Gram matrix
μ_m	weight assigned to the m^{th} data set or kernel matrix
m	index parameter for the number of data sets or kernel matrices, $m = 1, \ldots, M$
M	number of data sets or kernel matrices
b	bias term in the class of hyperplanes
w	normal vector of the decision hyperplane
α_i	Lagrange multiplier; support value
ξ_i	slack variable
e_i	error variable
ζ_i	weight assigned to the i^{th} sample to take class unbalancing into account
γ, c	regularization parameter
r	grid dimension for tuning parameter optimization; range of variable z

υ	number of folds in cross-validation
T	time point
MT_j	model based on microarray data at time point j
PT_j	model based on proteomics data at time point j
MG	model based on microarray and genomics data
C	continuous variable
D	discrete continuous variable
O	ordinal variable
N	nominal variable
ρ	correlation coefficient
ϵ	variable of the recurrent HMM
W	weighted adjacency matrix
D	degree matrix
d_l	degree of node l
L	Laplacian matrix
G	Moore-Penrose pseudoinverse $L^\dagger$ of the Laplacian matrix of a graph
P, Q	probability distribution of a variable
R	joint distribution for two variables
I	mutual information between two variables

Contents

Voorwoord i

Abstract v

Korte Inhoud vii

Acronyms ix

Symbols xiii

Contents xv

1 Introduction 1

1.1 Molecular Biology and Cancer 1

1.1.1 Macromolecules 2

1.1.2 Data Types 6

1.1.3 High-Throughput Technologies 7

1.1.4 Genetic Alterations 11

1.1.5 Cancer as a (epi)Genetic Disease 12

1.2 Overview of Realizations in Cancer Research 15

1.2.1 Hallmark Microarray Studies 15

1.2.2 Concerns and Criticisms 17

1.2.3 Clinical Trials 19
1.3 Clinical Decision Support 20
1.4 Machine Learning Techniques 23
1.4.1 Kernel Methods 23
1.4.2 Dimensionality Reduction 24
1.5 Main Contributions of this Thesis 25
1.6 Chapter-by-chapter Overview 27
1.7 Other Research 28
1.7.1 Microarray Studies 29
1.7.2 Proteomic Studies 30
1.7.3 Biomedical Studies 32

2 A Methodological Primer **35**

2.1 Evaluation Measures 35
2.2 Classification Methods 37
2.2.1 Kernel Methods 38
2.2.2 Support Vector Machines 39
2.2.3 Least Squares Support Vector Machines 41
2.2.4 Weighted LS-SVM 43
2.3 Feature Selection 44
2.3.1 Univariate Feature Selection 45
2.3.2 Differential Expression via Distance Synthesis 45
2.3.3 Optimization of the Feature Subset 46
2.4 Parameter Setting Methods 46
2.4.1 Error Estimation Methods 46
2.4.2 Grid Search 47
2.5 Conclusion 48

3 High-Dimensional Kernel-Based Data Integration 49

3.1 Introduction 49

3.2 Review on Data Integration 51

3.3 Data 52

3.3.1 Data Set I on Rectal Cancer 53

3.3.2 Data Set II on Prostate Cancer 55

3.4 Methodology 57

3.4.1 Data Fusion 57

3.4.2 Model Building 59

3.4.3 Weight Optimization 61

3.5 Toolbox HI-DID-IT 62

3.6 Results 64

3.6.1 Study I on Rectal Cancer 64

3.6.2 Study II on Prostate Cancer 70

3.6.3 Comparison with an Ensemble Method 74

3.6.4 Generalizability Analysis 74

3.6.5 Correlation Analysis 76

3.7 Conclusion 76

4 New Kernel Function for Modeling Clinical Data 79

4.1 Introduction 79

4.2 Data 80

4.2.1 Clinical Data Sets 81

4.2.2 Clinical and Microarray Data 85

4.3 Methodology 87

4.3.1 Kernel Function for Clinical Data 87

4.3.2 Model Building 91

4.4 Results 92

4.4.1 Comparison of the Linear and Clinical Kernel Function . . 92
4.4.2 Results for the Clinical Data Sets 95
4.4.3 Results for Data Integration 96
4.4.4 Importance of the Clinical Kernel in a Broader Spectrum . 98
4.5 Conclusion 102

5 Modeling Array CGH Data with HMM-Based Feature Selection 103
5.1 Introduction 103
5.2 Data 105
5.3 Methodology 106
5.3.1 Recurrent HMM 106
5.3.2 Model Building 107
5.3.3 Functional Annotation Analysis 107
5.4 Results 109
5.5 Conclusion 113

6 Improved Decision Support with Graph Encoded Interactome Data 115
6.1 Introduction 116
6.2 Data 119
6.2.1 Microarray Data Sets 119
6.2.2 Secondary Data Sources 121
6.3 Methodology 128
6.3.1 Pseudoinverse Laplacian 128
6.3.2 Adapted Kernel Function 130
6.3.3 Combining Classifiers 130
6.3.4 Model Building 131
6.4 Results 132
6.4.1 Selection Outstanding Secondary Data Sources 134

6.4.2 Combination Multiple Classifiers 135
6.4.3 Training and Test Results 136
6.5 Conclusion 142

7 Conclusions and Future Research **145**
7.1 Conclusions 145
7.2 Future Research 148
7.2.1 Medical and Technological Trends 148
7.2.2 Methodological Challenges 149
7.2.3 The Future of Clinical Decision Support 151

A ROC Curves **153**

B Generalizability Results **159**

Bibliography **163**

Curriculum vitae **199**

Publications by the author **201**

Chapter 1

Introduction

Biomedical research is becoming a cross-disciplinary endeavor. Due to diverse activities required for diagnosis and management of cancer (for example, sample collection, data analysis and development of new diagnostic approaches), a strong collaboration is needed among biologists, engineers, physicists and mathematicians. We want to contribute towards a shift in approach from traditional clinicopathological markers - properly working for a similar population of patients - to personalized clinical decision support with a genomic basis for diagnosis, prognosis and treatment decisions in individual patients.

1.1 Molecular Biology and Cancer

Cancer is one of the most common and severe diseases seen in clinical medicine. In Europe, it is an important public health problem with an estimated 3.2 million new cases each year, primarily of breast, prostate, colorectal and lung cancer. The annual number of deaths equals 1.7 million, mainly caused by lung, colorectal and breast cancer [117]. The ageing of the population will further increase these numbers. Moreover, cancer is responsible for more than 10% of the total cost of medical care. As cancer is fatal when not treated, early diagnosis and treatment are of vital importance. Before going into detail in cancer and its development, a short primer on molecular biology is given, with an overview of the three most important types of macromolecules in the cell (DNA, RNA and the protein), the genetic flow of information known as the *central dogma*, and the different biological data types with corresponding technology. We end this section with genetic alterations that

underlie developmental abnormalities and diseases, especially cancer. We refer to [6] and Chapters 3 and 16 in [238] for a more detailed description.

1.1.1 Macromolecules

The hereditary information needed to specify all aspects of embryogenesis, development, growth, metabolism and reproduction in an organism is stored in each nucleated cell in the form of *deoxyribonucleic acid (DNA)* molecules, referred to as the *genome* and studied in *genomics*. DNA is a linear chain composed of four basic molecules called *nucleotides*. Each nucleotide contains phosphate, sugar and one of the four bases adenine (A), cytosine (C), guanine (G) and thymine (T). Due to the complementarity between bases A and T and bases G and C, the same base pairs are always formed. As can be seen in Figure 1.1, the native state of genomic DNA is a double helix of two complementary strands, tightly packed and organized in each cell. The binding of two complementary DNA molecules is referred to as *hybridization*. The humane genome consists of 3.2×10^9 nucleotides, separated into 23 chromosomes and containing about 25,000 *genes* as subsequences. A gene is defined as a DNA segment that is required for the production of a functional product. The coding regions, called *exons*, are interrupted by non-coding regions called *introns*. In 1990, the Human Genome Project was launched in the United States to obtain the complete sequence of 3.2 billion base pairs, to create detailed maps for all chromosomes, and to identify the 25,000 human genes [258]. In February 2001, drafts of the human genome were published in Nature [68] and Science [349]. In April 2003, the Human Genome project was announced to be completed, with more than 99% of the human genome being sequenced [269]. Both studies in 2001 reported that exons only constitute approximately 1% of the human genome, that 25% of the genome is composed of intronic sequences, and 75% represents intergenic DNA for which the function is not yet known. Finally, the genome of any two individuals differs by only 0.1%.

Ribonucleic acid (RNA) is a single-stranded chain of ribonucleotides, slightly differing from nucleotides with thymine replaced by uracil (U). Genomic DNA directs the synthesis and sequence of RNA through a process known as *transcription* (see Figure 1.1). When a protein-coding gene is transcribed, the RNA strand is referred to as pre-messenger RNA and includes all introns and exons. This strand is then converted and shortened to a *messenger RNA* (mRNA) molecule, by removing the introns and sealing the segments corresponding to exons in a process called *splicing*. The collection of all transcribed genes is referred to as the *transcriptome*. *Transcriptomics* focuses on patterns in gene expression to decipher the underlying biology. Besides mRNA, other non-protein coding parts of the genome are transcribed into RNA molecules. Transfer RNA and ribosomal RNA molecules are essential for translation (see next paragraph), while others play structural roles in various nuclear and cytoplasmic processes.

MicroRNAs (miRNAs) are small, highly-conserved non-coding RNAs and control gene expression post-transcriptionally. They act as key elements in cellular differentiation and participate in embryonic development [284]. Approximately 1,000 miRNAs in the human genome are believed to regulate the expression of on average 200 mRNA targets, either by inducing degradation of the target or by blocking protein translation without degrading the targets. In total, $> 30\%$ of the human genome may be under translational regulation by miRNAs.

While the latter RNA molecules are final gene products, mRNA is used as a blueprint for the synthesis of *proteins*, the major building blocks of the cell. During the process of *translation*, mRNA is read in consecutive groups of three nucleotides, constituting a *codon* that is specific for a particular *amino acid*. Proteins are thus linear polymers composed of 20 different types of amino acids. As can be seen in Figure 1.1, proteins fold and bind into a specific three-dimensional structure, assumed to be uniquely determined by the sequence of amino acids. A *domain* is a subunit of a protein that folds separately into its native structure. The structure of a protein and the physicochemical properties of the amino acids exposed on the protein surface determine the molecules the protein can bind to and its functionality.

An important mechanism is that of *alternative splicing*. As said earlier, only exons are retained in mRNA. However, certain proteins can cause exons to be lengthened, shortened or even completely skipped. The same gene can thus give rise to multiple related but different mRNAs, with the pattern of splicing, for example, depending on the tissue in which transcription occurs. At least one third of all human genes undergo alternative splicing, with on average two or three alternative transcripts per gene. After translation, proteins can also undergo *post-translational modifications*. The activity of many proteins is, for example, regulated by phosphorylation or acetylation, altered by chemical modifications of amino acid residues (methylation), or processed by cleavage in smaller polypeptide chains [63]. These two principles of alternative splicing and post-translational modifications are represented in Figure 1.1 and contribute to the overall protein diversity in each organism. The 25,000 human genes can encode a million distinguishable functional entities at the protein level, with on average 8 to 9 distinct proteins per gene. The collection of all proteins, referred to as the *proteome*, is therefore many orders of magnitude larger than the transcriptome and reflects the functional state of the cell much more accurately. This makes the proteome a potential richer source of data for unraveling diseases and gathering dynamical information [186]. How, when and where proteins are expressed, their concentration, modifications, protein-protein interactions and regulation are studied with *proteomics* [152].

In order to understand the physiology of a cell or organism, not only genes and proteins need to be studied, but also products of the action of proteins, being *metabolites* [236] [271]. Metabolites such as metabolic intermediates, hormones and other signaling molecules are involved in normal growth, development

and reproduction. The *metabolome* refers to the complete set of small-molecule metabolites formed by the cell in association with its metabolism, and *metabolomics* is the study of all metabolites within a cell or tissue. As represented in Figure 1.2, after the genome, transcriptome and proteome, the metabolome can be regarded as the end point in the era of individual -ome levels with metabolites closely linked to the phenotype of an organism. Metabolomics can therefore give an instantaneous snapshot of the physiology of the cell. It provides a more comprehensive view of how cells function to increase the understanding of the underlying biochemical processes that are altered in disease states. Contrary to the transcriptome and proteome, however, metabolites are much more complex to study. Several 10,000 metabolites are known so far, with many still unknown. These molecules are very heterogeneous with distinct chemical properties and large differences in concentration. Finally, databases on metabolites are missing or incomplete, and a link with the genome lacks.

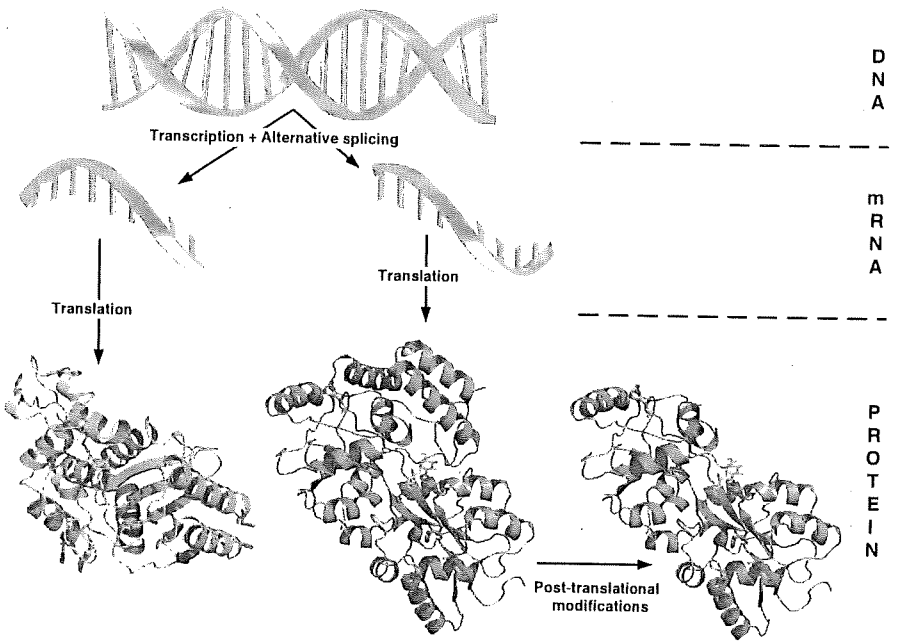


Figure 1.1: *Overview of the central dogma of molecular biology:* DNA segments in the genome that correspond to genes are transcribed to messenger RNA molecules. In case of alternative splicing, one gene results in multiple distinct mRNA molecules. Transcription is followed by the process of translation in which mRNA is used as template for protein synthesis. Proteins can further undergo post-translational modifications, increasing the overall protein diversity.

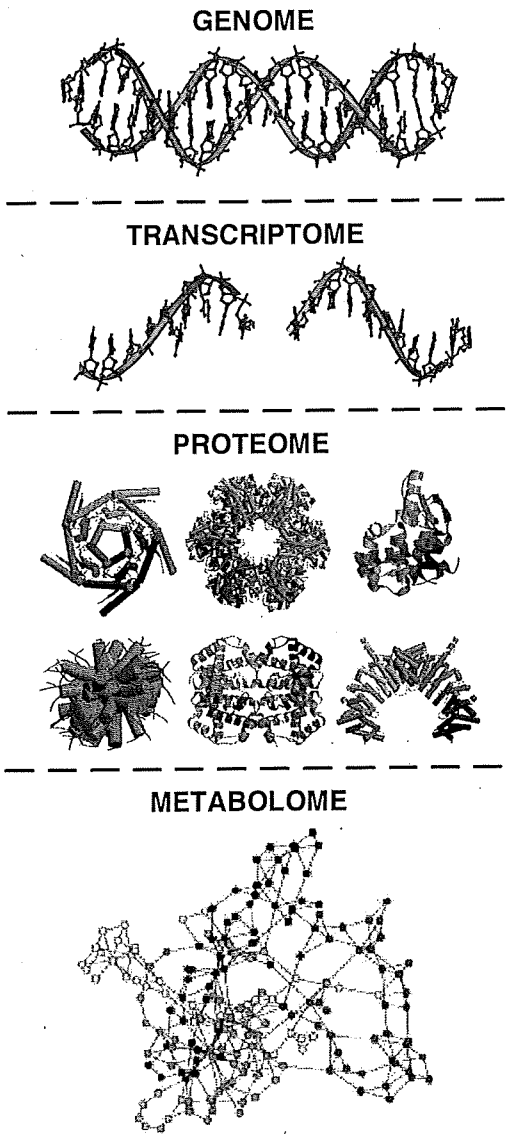


Figure 1.2: *Multiple -ome levels of biological regulation at which information can be gathered:* the genome stores hereditary information in the form of DNA; the transcriptome is the collection of all transcribed genes; the proteome refers to all proteins; the metabolome refers to all small-molecule metabolites.

1.1.2 Data Types

With the previous macromolecules, large volumes of expression, interaction, sequence and organism data can be generated on various aspects of the molecular machinery in cells (see Section 1.1.3 for an overview of the technologies) [223]. For expression data, a distinction is made between the abundance of mRNAs (gene expression), the abundance of proteins (protein expression) and the concentration of metabolites (metabolite expression). Interactions can be defined on several levels: between proteins to form complexes, between proteins and metabolites (for example, a receptor-ligand binding), and between proteins and DNA to, for example, initiate transcription. In Chapter 6, an overview is given of databases on interactions, used for the construction of biological networks with nodes representing genes or proteins and edges representing interactions between these entities. Sequence data, although not investigated in this dissertation, are represented as a string of nucleotides for DNA, of ribonucleotides for mRNA, and of amino acids for proteins. For data at the organism level, a distinction is made between the *genotype* referring to all heritable information giving rise to an organism, the *phenotype* referring to the observable physical properties of the organism such as size, gender and eye color, and broader clinical data on disease status. Finally, information is often accessible on environmental factors, such as fat and alcohol consumption, smoking, excess exposure to sunlight, chemicals and viruses.

Despite this wealth of data types, the main focus for clinical decision support has been on gene expression data. This data type, however, only represents one view on how cells and tissues are characterized [36]. As certain events associated with malignant phenotypes are reflected on the DNA and protein level rather than on the transcriptional level, combining data from multiple -ome levels is particularly appealing in order to comprehensively profile transcriptional, translational and post-translational programs in cells and tissues. Besides the established genome, transcriptome and proteome, also other -ome levels are emerging. The metabolome, on one hand, has been introduced in Section 1.1.1. On the other hand, in ~200 different cell types in the human body, distinct gene expression patterns at numerous developmental stages are needed. An additional layer of information is therefore required that exceeds the information of the four bases in the genome [305]. Proper cell- and tissue-specific gene expression is achieved with epigenetic modifications, known as the *epigenome*. These modifications do not alter the DNA sequence but are heritable and involved in gene transcription regulation and DNA repair. *Epigenomics* is the study of epigenetic changes such as DNA methylation, modifications in chromatin structure and condensation, and changes involving RNAs that maintain the gene transcription state in a heritable manner [114] [158] [252]. *DNA methylation* is the addition of a methyl group to the cytosine base in a CpG dinucleotide. This process is essential during embryonic development. Aberrant DNA methylation, however, has been

shown to be involved in diseases when compared to the normal genome. Loss of methylation at normally methylated sequences is referred to as *hypomethylation*, while *hypermethylation* is defined as the gain of methylated sequences at sites usually unmethylated [252].

Also -omics fields covering the individual -ome levels are arising. *Functional genomics* interprets DNA in terms of genes to decipher the underlying biology by studying the functionality of specific genes, their associated proteins, the biological processes in which they participate and the diseases they are related to. *Systems biology* goes beyond functional genomics because the information flow hierarchically continues beyond the central dogma with protein interactions, pathways, cells, tissues and organisms. It gathers information from each of these distinct levels to investigate the relationship between all elements in a particular biological system as part of the system's response to environmental or genetic perturbation [167]. *Interactomics* encompasses the physical interactions between all molecules within a cell to understand how complex networks of interactions impact on phenotype and how dynamic changes in these networks can lead to human diseases [351]. *Fluxomics*, finally, deals with dynamic changes of molecules within a cell over time.

1.1.3 High-Throughput Technologies

For most data types discussed in the previous section, more than one measurement technology exists with the potential to supply a wealth of predictive molecular markers. An overview of these technologies is provided in this section.

Transcriptomics

Multiple technologies are available to simultaneously measure the expression of a large number of genes in a biological specimen. *Microarray technology* is the most popular technology for large-scale expression measurements of almost all known human genes in a single experiment [60] [259] [283]. A microarray chip contains highly specific sequences of individual known genes (that is, *probes*), complementary to the mRNA material in a sample of interest and located at unique positions on the array surface. Pools of mRNA are first reverse-transcribed to cDNA, labeled, and subsequently hybridized to the probes immobilized on the chip. After washing the array to remove unbound and nonspecific material, the amount of fluorescence emitted at each probe spot position is measured. Because the number of cDNAs from the sample that hybridize to one specific probe is approximately proportional to the corresponding mRNA concentration, the amount of fluorescence represents the abundance of a gene transcript. Depending on the chip design, gene abundance is represented as an absolute intensity value

for single-channel arrays or as an intensity ratio for two-channel arrays. For the latter, two spectrally distinct fluorescent dyes are used for the sample of interest and the reference sample, both hybridized to the same chip. Another distinction between single- and two-channel arrays is the use of a set of short probes, referred to as *probe set*, to represent the sequence of a gene in the single-channel array. In this dissertation, the expression of genes was studied rather than that of probes or probe sets in order to make data sets comparable and to get rid of dependency on platform. However, microarray data are high dimensional and noisy due to, for example, incomplete washing and nonspecific hybridization. Moreover, data from different samples cannot directly be compared due to varying hybridization efficiency and amounts of mRNA per cell. Both issues require data filtering, normalization and background correction [169] [372]. Microarray technology can not measure the complete transcription profile either due to a limited number of probes per gene on a chip, nor does the transcriptome completely reflect the biology underlying a disease.

Other technologies for gene expression are *sequencing-based methods* that do not require prior gene knowledge, *low-density arrays* for a limited number of genes of interest, and *quantitative polymerase chain reaction (qPCR)*, which allows a very precise determination of expression levels for a limited number of genes with known sequence [60]. The principle of gene expression arrays and real-time PCR can also be used to measure miRNA expression profiles. With *exon arrays*, the expression of virtually all known and predicted exons present in the human genome can be measured [123] [193]. This allows the identification of disease-specific splice variants that are due to exon skipping or alternative splice site use.

Proteomics

Proteomic studies are technically more challenging than transcriptomic studies due to post-translational modifications, multi-protein complexes and the storage of proteins in distinct compartments. With *mass spectrometry (MS)*, peptides, proteins and polynucleotides are first ionized. A mass analyzer then separates the protein fragments based on their mass-to-charge (m/z) ratio. The m/z values of the fragments are derived very precisely from their time of flight after a certain acceleration, and the corresponding intensities are obtained with an ion detector [3] [60] [146]. The most commonly used MS techniques in translational research are MALDI (matrix-assisted laser desorption and ionization), SELDI (surface-enhanced laser desorption and ionization) and LC-ESI MS/MS (liquid chromatography coupled to electrospray ionization tandem MS). LC is a fractionation technique, while tandem MS leads to a more detailed fragmentation of the spectra than possible with a single mass spectrometer. These techniques can be used in combination with different mass analyzers such as time-of-flight (TOF) and quadrupole time-of-flight (QqTOF). *2D-gel electrophoresis*, another

fractionation technique, quantitates proteins by spatially separating them in a gel according to their mass and electric charge. Although these techniques are most popular, others exist such as *tissue, protein and antibody microarrays* [60] [216]. *Oncoproteomics*, which is the study of proteins and their interactions in cancer cells, provides hope that novel biomarkers will be discovered with the advent of new and improved proteomic technologies [58].

Genomics

The collection of genes considered with genomics is not limited to protein-coding genes. Also other elements with important functions in the genome are studied, such as regions encoding miRNAs, transcription factor binding sites and evolutionary conserved regions. Regions of increased or decreased DNA copy number can be detected using the *array comparative genomic hybridization (CGH) technology*. This high-throughput technique measures variations in DNA copy number (CNVs) within the entire genome of a disease sample compared to a normal sample [250]. In an array CGH experiment, total genomic DNA from tumor and normal reference cell populations are isolated and subsequently labeled with different fluorescent dyes before being hybridized to several thousands of probes on a glass slide (see Figure 1.3). This allows calculating the log ratios of the fluorescence intensities of the tumor to that of the normal reference DNA. Because the reference cell population is normal, an increase or decrease in log intensity ratio indicates a DNA copy number variation in the genome of the tumor cells, such that negative log ratios correspond to deletions (losses), positive log ratios to gains or amplifications and zero log ratios to neutral regions in which no change occurred. CNVs within regions of at least 5-10 kilobases in length can be detected. Many small aberrations have emerged as prognostic and predictive markers. Numerous aberrations, however, also affect large genomic regions, encompassing multiple genes or whole chromosome arms.

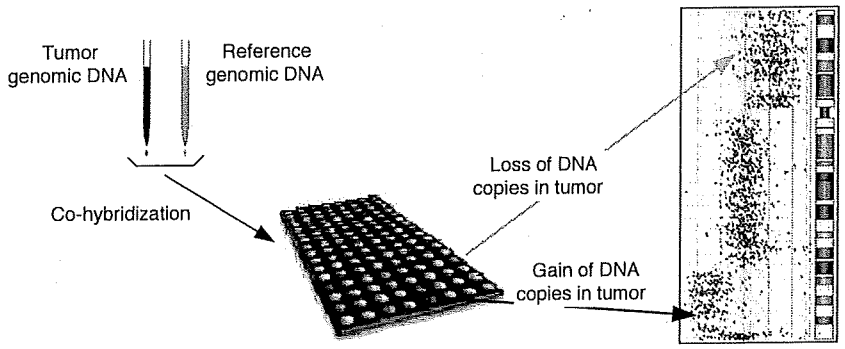


Figure 1.3: *Principle of the array CGH technology*

Another technique used in genomics is *single nucleotide polymorphisms (SNP) genotyping*. This approach allows to simultaneously study at high resolution several types of genetic changes, such as CNVs for the identification of novel oncogenes and tumor suppressor genes, and loss of heterozygosity (LOH) events, which are somatic conversions of heterozygosity to homozygosity that can lead to the inactivation of tumor suppressor genes [70] [106] [383]. These dense genotyping chips make genome-wide association studies on thousands of cases and controls technically and financially feasible, leading to the identification of genes involved in common human diseases [44].

Next-generation genome sequencers provide accurate, genome-wide and unbiased quantitative and qualitative measurements for both DNA and RNA, by measuring the sequence of nucleotides for all genes in a tumor cell [201] [295]. It is therefore applicable in a wide range of research applications, such as mRNA and miRNA expression, protein-DNA interactions, genetic variation profiling, and chemically modified DNA [177]. The improvement in sequencing technology has been crucial in deciphering the first human reference genome [315]. However, it still took 13 years to complete the first human genome with a cost of three billion dollars due to the requirement to introduce DNA pieces into bacteria. A further evolution from first- to second-generation sequencers, such as SOLiD [296] and Solexa [22] reduced both cost and time to one million dollars and two months of work by twenty people to decipher the genome of James D. Watson, one of the discoverers of the structure of DNA in 1953 [361]. Comparison of a set of genes in a large amount of samples is now affordable due to a five- to ten-fold decrease in cost. With the introduction of third-generation sequencers, the Personal Genome Project [61] may become feasible when 3 billion of base pairs can accurately be sequenced in several hours for thousand dollars or less (see also http://www.1000genomes.org/).

Other Technologies

In newer -omics fields, previous described technologies are used or newer techniques have been developed. Concerning metabolomics, sophisticated analytical technologies are required for the identification and quantification of cellular metabolites. *Mass spectrometry* coupled to different chromatographic separation techniques such as liquid or gas chromatography can be considered as the most important analytical approach in metabolomics studies [146] [271]. It provides high sensitivity and mass accuracy and allows real-time metabolite identification. However, not all compounds form ions and the structure of unknown compounds can hardly be determined with mass spectrometers. Despite of orders of magnitude less sensitive, *nuclear magnetic resonance (NMR) spectroscopy* is used in metabolomics ascribed to its reproducibility and ability to quantitate compounds and identify unknown metabolites. No single technology, however, allows the analysis of the

entire metabolome due to the huge diversity of chemical structures and the large differences in abundance.

For epigenomics, DNA methylation can be detected in blood, sputum and biopsy specimens. Novel genome-wide DNA methylation markers can therefore be discovered with high-throughput approaches such as microarray expression profiling before and after pharmacological inhibition of DNA methylation [291]. Results obtained with global approaches, however, need to be independently confirmed by other methods. A low-cost method with higher resolution is *quantitative multiplex methylation-specific PCR*, which determines the relative percentage of methylated and unmethylated DNA in a sample with two rounds of PCR amplification [115]. More recently, mass spectrometry was used to quantitatively assess methylation at given CpG sites [108]. Finally, *tiling arrays* can be used more widely for functional genomics, in order to detect epigenetic modifications, splice variants and altered gene expression and regulation. All these approaches and technologies should be considered as they promise to be useful in unraveling cancer mechanisms, discovering cancer-causing genes, and the refinement of their molecular descriptions.

1.1.4 Genetic Alterations

Genetic variation in the human genome can take many forms and has been found in the general population as well as in all major tumor types [127] [250] [266]. In the human body, cells are constantly divided into two daughter cells with an estimated 10^{15} cell divisions during an adult's lifetime. Cell division also requires duplication of the heritable information encoded in the genome. However, replication errors occur with a frequency of one in 10^{10} bases. Such modifications of the heritable information in reproductive cells are referred to as *mutations*. More specifically, a *point mutation* is an erroneous change of a single nucleotide in the genome. Most of these mutations are detected and corrected due to redundancy in the double helix structure. When not corrected, a point mutation may result in a change in gene expression or altered protein level, although the majority of the thousands of new DNA mutations in the genome of every cell do not have phenotypic effects. When different genotypes coexist in a population, these mutations are referred to as *SNPs*. There exist ~10 million SNPs throughout the genome, with on average one SNP per 400-1,000 base pairs [106]. They cause no discernable phenotype, but may predispose to disease by affecting gene function without altering the protein structure. SNPs may therefore determine response to therapy and susceptibility to many diseases.

Concerning structural chromosomal rearrangements, a distinction can be made between unbalanced rearrangements in which material has been added or is missing, and balanced rearrangements with the normal amount of chromosomal

material (see Figure 1.4) [127]. *Duplication* or *amplification* refers to one or many additional copies of a segment in the genome, enhancing the activity of structurally normal genes in the affected chromosomal region. Loss of a chromosome segment is referred to as *deletion* and reduces the expression level of the included genes, possibly disabling their normal function. These unbalanced rearrangements are collectively termed as *copy number variations (CNVs)*, a frequent, dynamic and complex form of genetic diversity that makes chromosomes partial trisomic or monosomic. The abnormality of the phenotype depends on the number and function of affected genes. In the general population, CNVs have been shown to cover 12% of the genome [266]. The larger fraction of structural changes in the genome compared to SNPs underscores the importance of CNVs in genetic variation [297]. Moreover, genome-wide association studies have resulted in many SNPs that are associated with a disease; however, without the ability to explain much of the disease's heritability. It is assumed that CNVs will explain a larger portion of the genetic basis of a disease [297]. Many laboratories have therefore shifted their approach from genocentric to genomic studies, to reveal CNVs and genomic rearrangements [214].

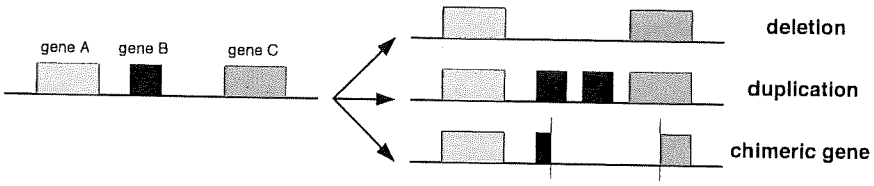


Figure 1.4: *Overview of structural chromosomal rearrangements:* Unbalanced rearrangements are deletions and duplications with loss and gain of a chromosome segment, respectively. Balanced rearrangements (inversion and translocation) can cause fusion of genes.

Balanced rearrangements do not usually influence the phenotype. An *inversion* occurs when a single chromosome undergoes two breaks, followed by reconstitution with the inverted segment between the breaks. *Translocation* involves the exchange of chromosome segments between two chromosomes. Both inversion and translocation can result in the formation of a chimeric gene with a new or altered function through the fusion of parts of two genes. Finally, gene promoters can be methylated or demethylated, thereby decreasing or increasing expression level.

1.1.5 Cancer as a (epi)Genetic Disease

Cancer is fundamentally a genetic disease, practically always associated with one or more chromosomal abnormalities [353]. It is initiated by mutations in

one or more genes that regulate cell growth and programmed cell death. These cancer-causing mutations can be inherited from one of the parents. Most cancers, however, are sporadic, meaning that at a certain time in life a life-threatening mutation occurs in the DNA sequence of an oncogene or tumor suppressor gene in a single sporadic cell. Proto-oncogenes are normal cellular protein-coding genes that promote growth and survival of cells. A gain-of-function mutation in the coding sequence, a mutation in a regulatory element, or amplified regions containing extra copies of proto-oncogenes can turn them into *oncogenes*. Proto-oncogenes can also be activated by translocation, constituting a chimeric gene with new oncogenic properties or activating an oncogene by placing it downstream from the promoter of a tissue-specific or housekeeping gene. Oncogenes stimulate cell division and proliferation in an abnormal way, thereby contributing to the malignant character of cells. *Tumor suppressor genes*, on the other hand, directly control cell growth (*gatekeepers*) or protect the integrity of the genome by repairing mistakes made during DNA replication or induced by exposure to mutagens (*caretakers*). Loss-of-function through mutations resulting in a truncated protein, missense mutations at residues that are essential for gene activity, deletions, translocations, epigenetic silencing or changes in the splicing pattern in gatekeeper genes lead to uncontrolled cell accumulation. Inactivated caretaker genes can no longer prevent mutations from accumulation in proto-oncogenes or other tumor suppressor genes.

Not only individual genes, but also large-scale genomic gains and extensive genomic deletions affecting multiple genes are frequent in many types of tumors and are associated with cancer progression and prognosis [65] [127]. There is strong evidence that aberrant splice isoforms that lack certain exons [123] and key metabolites from many pathways [329] are involved in the initiation and progression of cancer. Many chromosomal regions that are recurrently deleted in cancer, however, lack protein-coding genes. Increasing evidence shows that alterations in the expression of miRNAs contribute to the pathogenesis of human malignancies [112] [284] [380]. These alterations can be caused by deletions, amplifications or mutations, epigenetic silencing and deregulation of transcription factors that target specific miRNAs. These miRNAs, known as *oncomiRs*, regulate the expression of tumor-associated genes. Overexpression of an oncomiR suppresses the expression of target tumor suppressor genes while loss of function results in overexpression of the target oncogenes. Due to the large set of target genes per miRNA, a mutation in the corresponding DNA sequence contributes to a genome-wide transcriptional deregulation. It has been shown by Zhang and colleagues [380] that genomic copy number loss and epigenetic silencing may account for the down-regulation of miRNAs in cancer. Also Lu and colleagues [212] observed a general down-regulation of miRNAs in tumors compared to normal tissues. They were able to successfully classify poorly differentiated tumors using miRNA expression profiles. The mRNA profiles were lacking this information when applied to the same samples. Due to their diagnostic, prognostic and therapeutic implications for cancer patients, miRNAs can be used as molecular markers.

Genetic abnormalities found in cancer do not provide the complete picture of genomic alterations. Cancer is not only a genetic but also an epigenetic disease. Locus-specific hypermethylation at CpG islands in the promoter of tumor suppressor genes is a common hallmark that inactivates these genes. As half of the tissue-specific genes contain a promoter CpG island, hypermethylation affects cellular pathways with a tumor-specific profile [113]. So far, nearly 150 genes with a tumor-specific methylation have been identified in primary tumor samples, of which an overview is given in [290]. Demethylation in proto-oncogenes, on the other hand, leads to gene activation. DNA methylation events occur early in tumorigenesis, which makes this modification a perfect target for early detection of malignant cells. Genome-wide hypomethylation usually targets non-coding regions, influences genome stability and substantially alters the overall chromatin structure. However, it has rarely been used as molecular marker due to poor sensitivity and cancer type specificity [158]. Finally, histone modifications and chromatin remodeling are also deregulated in cancer [158].

Once a cancer is initiated, additional genetic damage accumulates through mutations in tumor suppressor genes. As shown in Figure 1.5, the imbalance between cell proliferation and cell death leads to an abnormal accumulation of cells, finally resulting in a mass or tumor [238]. In case of uncontrolled cellular growth, the tumor can invade neighboring tissues or can spread to more distant sites, called metastasis. These tumors are referred to as *malignant* and classified by site, tissue type, histological appearance and degree of malignancy. Tumors that do not invade or metastasize are not cancerous and referred to as *benign*. The original somatic cell with the cancer-causing aberration thus serves as a stem cell, constituting multiple lineages of genetically unstable cells, varying in degree of malignancy and differing in the specific set of accumulated aberrations. As the whole process in Figure 1.5 leads to alterations in transcription, translation, post-translational modification and gene and protein function, cancer is the phenotypic end point of numerous genomic, cytogenetic and/or epigenomic alterations that have accumulated within cells [57]. This complicates the identification of the causal element in a malignant process. The Catalogue of Somatic Mutations in Cancer (COSMIC) lists over 100,000 frequent mutations in human genes, obtained by analyzing the coding sequences in the genome [121]. As only a limited number of gene mutations are biologically meaningful, a census of 418 cancer genes to date that are mutated and causally implicated in cancer development has been conducted based on specialist literature [130]. This list revealed that more than 1% of the human genome is involved in cancer pathogenesis. Finally, cancer is also characterized by a complex interplay between the tumor and surrounding normal tissue. The latter provides blood supply, permits cancer cells to metastasize to other organs and tissues, and protects the tumor from the immune system.

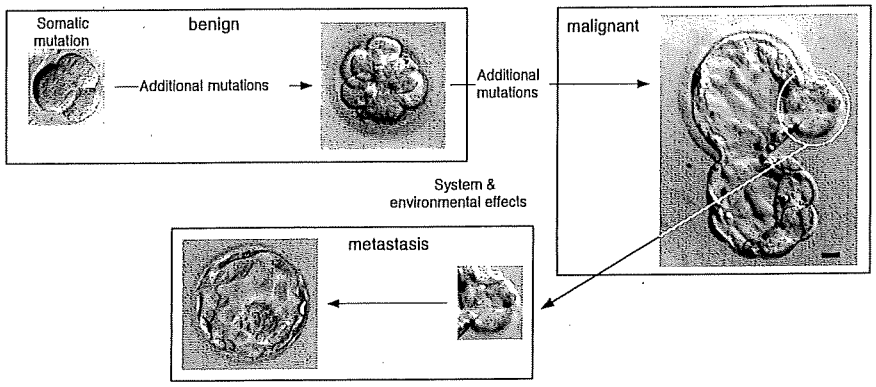


Figure 1.5: *Representation of the process of cancer progression with successive steps from mutation over primary tumor to metastasis [20].* As not all experimental results did fit this model, a parallel evolution model was suggested by Gray [145] in which cells from which metastasis develops do not derive from the primary tumor but are separated earlier during progression and evolve independently.

1.2 Overview of Realizations in Cancer Research

After an introduction on molecular biology, cancer, genome-wide data types and technologies, we will now provide an overview of the state-of-the-art research in cancer. Patient history and ultrasound parameters such as age, tumor size and histological grade, laboratory analyses of, for example, hormone receptor status and cancer antigen 125 (CA 125), as well as combinations thereof have been evaluated as prognostic indicators. Decisions concerning diagnosis, prognosis and therapy response are mainly based on these clinical and pathological indicators. However, patients with the same clinicopathological characteristics may have a markedly different phenotype and treatment response due to genomic and molecular differences. Traditional prognostic indicators are therefore not sufficient to capture the complexity of cancer and to tailor individualized therapy, necessitating the use of the technologies described in Section 1.1.3.

1.2.1 Hallmark Microarray Studies

Microarray technology has revolutionized basic biological science and the entire field of cancer taxonomy, biomarker development and identification of prognostic and predictive markers. As cancer is thought to be caused by genetic aberrations, microarrays were first used a decade ago to classify tissues as cancerous or non-cancerous [7] [143]. Within the domain of cancer, microarray technology has earned a prominent place for its capacity to characterize the underlying tumor

behavior in detail, with over the last years an enormous increase in the number of microarray studies on a wide range of cancer types [140]. Despite of the small number of tumors, these initial studies have illustrated that molecular profiling is a useful tool to improve our understanding of the complex biology and to reveal important genes and pathways that underlie the disease outcome.

Besides deciphering the biology, microarray technology has the potential to improve early diagnosis, refine cancer prognosis and revolutionize clinical management of cancer. For this purpose, molecular markers are combined in signatures to accurately predict the class membership of new samples. Two landmark studies in this field are those of van't Veer and colleagues [347] and Wang and colleagues [359], both applied to breast cancer. As many patients with breast cancer receive unnecessary treatment when based on conventional markers, the aim in both studies was to more accurately identify those patients who need further treatment while sparing a substantial proportion of women who are cured by only surgery. In the former study [347], a signature of 70 genes that correlated with survival was developed for the prediction of distant metastases within 5 years, based on a set of 98 lymph node negative patients under the age of 55. The results were internally confirmed on a set of 295 patients, inhomogeneous for lymph node status [332]. TRANSBIG, the TRANSlational research network of the Breast International Group, retrospectively validated this signature on an independent series of 302 patients from 5 different European centers, using an independent laboratory for the experiments and an independent statistical facility [45]. Finally, a prospective study on 427 patients showed that collection of good-quality tumor tissues and use of the gene signature is feasible in Dutch community hospitals [42]. A similar set-up in the second study [359] resulted in a 76-gene signature for the prediction of distant tumor recurrence. Analyses were based on 286 lymph node negative patients from all age groups with variable tumor size. This signature was validated on 180 independent samples from four European centers [120], as well as on the series from the TRANSBIG consortium [98]. The independent multicentric validation of both signatures proved strong robustness and reproducibility when adherence to a standard protocol was maintained, and increased evidence of clinical relevance as well. Moreover, both signatures performed similarly [98], possibly due to proliferation as common driving force [364]. A study from a systems-level perspective on 24 independent signatures with slight overlap also suggested that these signatures captured a common cancer cell phenotype [307]. Evidence for a significant association at the genome, transcriptome and proteome level followed from common transcriptional programs on cell proliferation and death, functional and molecular associations, and more closely related proteins in the interactome network than expected by chance.

1.2.2 Concerns and Criticisms

However, diagnosis and prognosis of cancer with microarray technology have been subject to controversy. The development of the 70-gene signature involved several critical issues. Samples used for model building and internal validation partially overlapped, and mainly younger patients with stage I and II breast cancer were included so that this signature may not be optimal for patients with other characteristics [149]. The predictive success of both studies was also criticized as the 70- and 76-gene signatures had only three genes in common. This disparity is not only attributed to differences in technology, sample preparation, mRNA extraction and patient characteristics. Ein-Dor and colleagues [109] showed that gene signatures are not unique. Using the same patients for training as in the original study [347], seven sets of 70 lower-ranked genes were found with the same prognostic capability as the published gene signature. Moreover, for more than 70 percent of 1,000 different splits of the data in a training and test part, at least one and on average four classifiers based on lower-ranked genes were capable of predicting survival with the same or better performance than the one based on the 70 top-ranked genes. These results show that gene selection strongly depends on the samples used for training, ascribed to the large amount of genes with similar correlation to outcome [189]. Prognostic signatures derived from one cohort also performed very differently in other cohorts [294]. These results raise doubt about reliability, robustness and stability of reported predictive gene lists. Michiels and colleagues [227] reanalyzed data from seven large microarray studies addressing cancer prognosis in six different cancer types. The originally reported prediction performances seemed overoptimistic, with five of the seven studies not classifying patients better than by chance when repeating a random training-validation strategy a sufficient number of times. These results emphasized instability of estimations based on a single validation set. This pessimism, however, was recently suppressed by Fan and colleagues [116]. Better generalization results were obtained when using other common feature selection methods, classifiers and validation strategies.

Serious issues concerning inappropriate statistical analysis and reporting of microarray results also need to be resolved before this technique can be translated into routine clinical use [303]. Despite clear guidelines on gene signature development and validation [302], most signatures to date violate at least one of the required criteria [140]. Firstly, data overfitting occurs in statistically underpowered studies, leading to overly optimistic training results and a poor performance on independent data. Secondly, in most biomedical studies cross-validation (CV) procedures need to be applied due to the limited number of samples (see Section 2.4.1). This requires gene selection and classifier construction to be properly repeated within each CV iteration. Ambroise and McLachlan [9], however, showed that in several published studies gene selection occurred prior to model building with CV. In the study of van't Veer and colleagues [347], both

a correct CV approach and an incomplete CV method without the reselection of differentially expressed genes were used. The invalid, smaller error estimate, however, received more attention. Next, the gain in predictive accuracy obtained with a gene signature should be compared with established clinical prognostic factors. According to Dunkler and colleagues [104], at least 20% of the variation in outcome should be additionally explainable by the gene signature on top of clinical characteristics before genes have a sufficient impact on an individual patient. Adding the 70-gene signature to age, nodal involvement, ER status and histological grade, however, could only increase the explained variation with 3%, suggesting that this signature does not perform noticeably better than indices based on different subsets of conventional markers [107]. Subsequently, due to the noisy characteristics of microarray data and the increase in analytical possibilities, discoveries can only be trusted when similar results are obtained with the same method in different data sets [168]. Moreover, due to the limited number of samples, gene selection suffers from the problem of multiple testing. P-values obtained with a statistical test and reflecting how genes distinguish two groups of samples can only be used for gene ranking. When developing predictive gene signatures, the number of false positive differentially expressed genes should be adequately controlled [105]. Finally, Baggerly and colleagues [13] examined five case studies in which microarray-based signatures of drug sensitivity were derived from cell lines to predict patient response, and concluded that common errors and irreproducibility may put patients at risk.

The above concerns and criticisms do not only apply to microarray data. In the last decade, mass spectrometry techniques were also used for classification purposes; however, with poor reproducibility of the results between laboratories [363]. In one of the earlier proteomic studies by Petricoin and colleagues [247], proteomic patterns in serum that completely discriminated 50 ovarian cancer patients from 50 unaffected women were identified with SELDI-TOF. When applied to an independent set of 116 serum samples, their discriminatory pattern yielded a sensitivity of 100% and specificity of 95%. As the first experiment was run with a low-resolution mass spectrometer with unacceptable week-to-week and machine-to-machine variability, high-resolution MS was subsequently used for the same data set (experiment 2) and a new one (experiment 3), yielding a 100% sensitive and specific model for ovarian cancer detection [67]. The SELDI spectra from these three experiments were reanalyzed by Baggerly and colleagues [14]. They found that results were not reproducible across experiments due to baseline variation, m/z values derived from one data set were not generalizable to the other data sets, and a major shift in protocol occurred during experiment 1. Moreover, perfect classification was only obtained with low-mass features in noisy regions, likely the result of instrument artifacts. These findings strengthen the evidence that the proteomic profiles published in [67] and [247] may be ascribed to artifacts in sample processing instead of the underlying cancer mechanisms. Careful experimental design and calibration are therefore essential to bring proteomics to the clinic.

1.2.3 Clinical Trials

Despite the above described controversies, the first gene expression profile signatures are being validated prospectively. These clinical trials are essential to obtain enough evidence for implementing gene arrays in daily practice for diagnosis and prognosis. The MINDACT trial (MIcroarray for Node Negative Disease may Avoid ChemoTherapy) [48] was launched by the Breast International Group to compare a standard clinicopathologic criterion with the 70-gene signature, available as the commercial laboratory test Mammaprint developed and marketed by Agilent. This 6,000-patient randomized, multicentric trial aims at proving that 5-year disease-free survival does not deteriorate when using the gene signature, while at the same time significantly reducing the percentage of women that is administered chemotherapy. The focus is therefore on those women for whom the clinicopathologic criterion and gene signature are discordant. Treatment is then randomly based on one of the two criteria. This trial is open in nine countries and 2,264 patients were enrolled after almost 3 years (http://www.eortc.be/services/unit/mindact/MINDACT_websiteii.asp). A second expression-profiling platform, the Oncotype DX assay based on 21 genes [241] [312], has been externally validated [151] [242] and is now being prospectively validated in the United States in the TAILORx trial (Trial Assigning IndividuaLized Options for treatment [Rx]) [240] [274] [311]. In this trial, all patients are administered hormonal therapy, but their benefit from chemotherapy is predicted. Chemotherapy is administered (or not) to patients with a high (or low) recurrence score obtained with the Oncotype DX test, while administration is randomized for those with intermediate scores. To date, 8,530 patients out of the 10,046 targeted women have been included (https://www.ctsu.org/public/).

Obstacles to the translational use of these signatures are expertise regarding RNA isolation and sample hybridization, and the requirement for fresh frozen tissue. Heterogeneity of the tumor sample is another issue due to the presence of necrosis, stroma and inflammatory infiltrates [149]. Variation in the histological content of biopsies has been shown to have an impact on expression profiles. In the study by Cleator and colleagues [64], almost 150 genes were found to positively or negatively correlate with the proportion of tumor cells from 43 patients. Consideration of biopsy composition would improve accuracy, as shown on subsets of samples that were more homogeneous in terms of cellular composition. The Oncotype DX assay would be a better suited clinical assay platform, applicable to formalin-fixed, paraffin-embedded tissue and using qRT-PCR to measure gene expression. Although these clinical trials are ongoing and no strong claims about the clinical value of the signatures can be made at this stage, the US Food and Drug Administration (FDA) has already approved some of the expression-profiling platforms for clinical use. In North America, the Oncotype DX assay has been widely accepted by the oncology community. Although preliminary results of the TAILORx trial showed that half of the accrued patients were at intermediate risk

and that the assay may be of no use [366], 7,500 physicians have already ordered the Oncotype DX assay to determine treatment for over 65,000 patients [140]. Moreover, as complete results from these large, randomized prospective trials will not be available for many years, it is possible that by the time of completion, this gene-based approach has become obsolete.

1.3 Clinical Decision Support

As follows from the previous section, more research in this area is definitely needed before microarray-based research discoveries can be used in a routine clinical setting where patients can directly benefit from them. *Clinical decision support (CDS)* systems are essential to improve cost-effectiveness of the overall health care system, especially since the cost of health care for age-related diseases such as cancer is significantly increasing due to the aging population. The general aim of decision support systems is to automate decisions based on domain knowledge and training data because problems become increasingly complex, humans are poorly equipped to deal with complexity, and they may have difficulties to objectively act under stress, time pressure and in high risk situations. Decision support improves reliability of decisions and reduces labor-intensive procedures by automatically classifying new data. In medical applications, there is a trend towards *clinical-aided systems* to assist clinicians in their decision making due to legal issues. A computer-based system thus offers a set of alternatives or executes a selection if the clinician approves [19]. This contributes towards the trend of 4P medicine - predictive, personalized, preventive and participatory - with increasing empowerment of the patient, driven by new technologies and the systems approach to medicine. The key challenge for health care, however, will be to provide affordable population-wide tools for common diseases, in order to improve the health of all people, rather than widening the gap between rich and poor [37].

A decade ago, CDS was defined as software designed to aid in clinical decision making, by matching characteristics of patients to a computerized knowledge base to obtain patient-specific recommendations for clinicians [166]. Such tools mainly enhanced clinical performance for drug dosing and preventive care, while the results for diagnostics were not convincing on a larger scale. CDS systems targeting diagnostic assistance are therefore urgently needed. In the field of cancer, the aim is to improve speed, accuracy and reliability of diagnostic and prognostic tools, pattern recognition, follow-up and screening compared to routine clinical management. Concerning therapy, CDS can help clinicians to improve the selection of patients who truly need therapy.

Clinicians are also overwhelmed by a *tsunami* of data for each single patient [223]. Not only clinical data on patient history, tumor characteristics, laboratory analysis,

ultrasound parameters, or environmental factors are accessible. Information can be unraveled from the genome, transcriptome, proteome and metabolome level, contributing to personalized medicine by improving quality of life. This also enables clinicians to make decisions regarding diagnosis, prognosis and response to therapy based on the actual biological reasons why subsets of tumors behave differently. Moreover, the amount of data generated by a variety of imaging modalities is increasing due to improved temporal and spatial resolution. These large volumes of data combined with factors such as time, cost and personalized treatment are complicating clinical decision making. Although clinicians prefer interpretable CDS systems [370], the shift towards genome-wide data requires non-traditional computational approaches to be introduced in clinical practice.

A nice illustration of a CDS system is the one developed by the International Ovarian Tumor Analysis (IOTA) group and represented in Figure 1.6 [324]. IOTA is a multicentric collaborative project for the pre-operative characterization of ovarian tumors as benign or malignant based on predictive computer models. The first phase of IOTA was conducted in 9 European centers between 1999 and 2002. 1,066 patients with an adnexal mass underwent a preoperative gray scale and color Doppler ultrasound examination. Information from these ultrasound images such as tumor size, locularity and papillary structure, and clinical data such as age, medical history and menopausal status were gathered according to a protocol and incorporated in several mathematical models to calculate the risk of malignancy. Between 2002 and 2005 three centers continued the prospective collection to internally validate the mathematical models developed in IOTA phase 1 [343]. The second phase of IOTA, conducted between 2005 and 2007, consisted of an external validation of the models in new centers with different population characteristics and different levels of ultrasound experience. The diagnostic algorithms were successfully validated prospectively on 1,938 patients with adnexal tumors in 19 centers throughout the world [325]. Currently, the best mathematical algorithm for the pre-operative classification of adnexal masses is prospectively tested in the third phase of the IOTA study by evaluating its prediction immediately after performing the ultrasound scan. This has been made possible by the development of a study screen that is included in the astraia software, used in most hospitals. Direct incorporation in a gynecologist's workflow fulfills the requirements for effective CDS such as speed, quality and usability [19]. Moreover, we have shown that the mathematical algorithm is generally applicable, independent of prevalence of malignancy [85]. The IOTA group has also confirmed in [324] that a CDS system needs to be well designed so that it allows less experienced clinicians to substantially improve their performance [369].

As a questionnaire in the UK showed that only 24 percent of 403 surveyed doctors would use a computer-based CDS system while 33 percent would never use one [369], efforts are needed to improve clinical acceptability. One possibility would be to combine a model's prediction with the expertise of a doctor [370]. We tested this

approach on the IOTA data because some adnexal masses are almost impossible to be classified confidently and correctly as benign or malignant, neither by an examiner nor by a CDS system based on ultrasound variables [88]. Second stage diagnostic tests are required to further investigate these masses in order to help the examiners. As shown in Figure 1.6, second stage tests can be pattern recognition in the hands of an experienced ultrasound examiner, additional approaches based on imaging such as ultrasound contrast agents and three-dimensional power Doppler ultrasound examination, or based on -omics data such as proteomic pattern analysis and tumor markers. In [88], we were able to identify the group of masses where characterization can be improved by second stage testing, and we investigated the minimal level of diagnostic performance of second stage tests required to improve overall diagnostic performance.

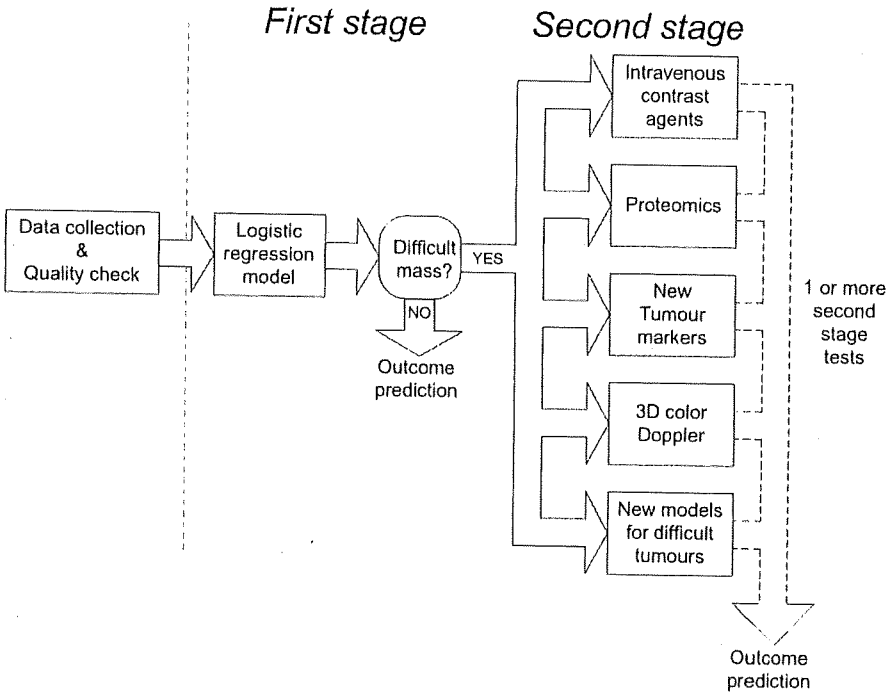


Figure 1.6: *Schematic overview of CDS in the IOTA study:* After performing an ultrasound scan, a logistic regression model is applied to immediately characterize the ovarian mass as benign or malignant. In case of uncertainty, second stage tests are required to further investigate the mass.

1.4 Machine Learning Techniques

Applied research is characterized by a single domain of interest, to find the best classifier or the best learning algorithm to apply in that domain. The aim in this dissertation, however, was to obtain learning strategies that work well over a wide range of application areas. *Machine learning (ML)* is a branch of artificial intelligence that induces patterns from past experience or large, complex data by optimizing a performance criterion [74] [188] [200]. Due to their good performance, clinical interpretability, ability to appropriately deal with missing and noisy data, and ability to reduce the number of attributes in the model, ML algorithms have extensively been used in cancer diagnosis and prognosis to transform the huge volume of data into biological knowledge. In [74], it has been shown that ML methods substantially improve performance with 15 to 25 percent compared to traditional statistical methods.

Three approaches can be distinguished for the design of a classifier [172]: firstly, samples can be assigned to a class, based on the concept of similarity (for example, nearest mean and 1-nearest neighbors); secondly, classifiers such as naïve Bayes and logistic regression are based on probabilistic principles; thirdly, decision boundaries obtained with, for example, perceptrons or support vector machines result from the optimization of an error-based criterion. As massive high dimensional data sets are becoming available in cancer research, we have chosen for the last approach in combination with dimensionality reduction.

1.4.1 Kernel Methods

The research group SISTA is characterized by a rich history on the use of kernel methods [8] [92] [245] [253] [377]. Those methods are a general and rich class of methods for pattern analysis [285]. In recent years, they have become a standard tool in data analysis, computational statistics, bioinformatics [27] and machine learning applications [292]. This rapid uptake is ascribed to their strong theoretical framework with numerical and statistical reliability, generalization-wise effectiveness and computational efficiency. In addition, they can handle high dimensional noisy data, extend linear methods to nonlinear settings, discover both linear and nonlinear relationships in data, and they have the capability to operate on a very wide range of data types. Kernel methods provide a vectorial representation of any kind of input items (for example, sequences, numeric vectors and molecular structures) by mapping them into a high dimensional space. In this space, data are represented as matrices without referring back to the original data and independent of the nature or complexity of the objects to be analyzed. The ability of kernel methods to deal with complexly structured data puts them in an ideal position for heterogeneous data integration.

A kernel algorithm for supervised classification is the Support Vector Machine (SVM) developed by Vapnik [348] and others. Contrary to most other classification methods and due to the way data are represented through kernels, SVMs can tackle high dimensional data. Even with a small number of samples, the SVM allows to train generalizable, nonlinear decision functions in a high dimensional space with maximum distance between the samples of two considered classes. Depending on the applied kernel function, this corresponds to a (non)linear discriminant boundary in the original input space. Because many biological problems involve high dimensional noisy data, SVMs have been applied very broadly in the field of computational biology, for example, for microarray gene expression analysis, prediction of protein-protein interactions, and peptide identification from MS data. We refer to Chapter 3 in [285] for a review on these and other pattern recognition problems.

In this dissertation, we used an extension of the SVM, the weighted Least Squares Support Vector Machine (wLS-SVM) [51] [318] [319]. The LS-SVM is easier and faster for high dimensional data because the quadratic programming problem for the SVM is converted into a set of linear equations. To account for the unbalance in many two-class problems, this linear problem has been extended with weights that are different for the positive and negative class of samples.

1.4.2 Dimensionality Reduction

High-throughput data suffer from the *curse of dimensionality* (that is, too many features) and the *curse of data set sparsity* (that is, too few samples) [308]. Due to the increase in large data sets, dimensionality reduction has become a necessity in many bioinformatics applications. It reduces the number of parameters that needs to be estimated from the data, increases the reliability of the parameter estimates for a fixed sample size, and it can alleviate the curse of dimensionality when the number of samples is limited. On the other hand, when many features are included, classifiers become more complex so that noise in the training set may be modeled. This phenomenon, known as *overfitting*, deteriorates generalization performance to new unseen data. The aim of feature selection is therefore two-fold: obtaining interpretable features with biological relevance, and a robust classifier that is insensitive to outliers with a high generalization power. Although dimensionality reduction introduces an additional layer of complexity, it is reasoned in biomedical applications due to redundancy in high dimensional data sets. Decisive features, however, are difficult to identify reliably. Jin [176] defined four key parameters in dimensionality reduction, that is, the number of samples, number of features, fraction of useful features and strength of useful features. He proved that a region in this four-dimensional parameter space exists where generalizable models can be obtained. However, prior knowledge of some useful features is essential for the identification of this region.

For dimensionality reduction techniques, a distinction can be made between feature extraction and feature selection methods. With feature extraction methods, the crucial information contained in the original data is transformed into a reduced set of new variables. This transformation is, for example, based on projection in case of principal component analysis. Although these transformed features may provide a better discriminative ability, they lack a clear physical interpretation. Feature selection methods, on the other hand, lead to reduced measurement costs as they do not alter the original representation of the variables. For biological interpretability, only feature selection methods were considered in this dissertation.

1.5 Main Contributions of this Thesis

In Section 1.3, evidence was given that the current tsunami of data emphasizes the need for more advanced statistical methods to transform the huge amount of data into meaningful support for clinicians. This dissertation is therefore directed towards analytical tools to study the heterogeneous disease, cancer. The focus is on translational research to move from bench to bedside and to introduce decision support systems in daily clinical practice. Translational research requires large interdisciplinary efforts in order to be successful, with an amalgamation of expertise and insights from cancer biology, cancer genetics, cancer modeling and clinical experiences. Computer scientists are needed for handling large data sets, mathematicians and engineers for data analyses, molecular biologists and physicists for the interpretation of the results and pathologists for microdissection [374]. A systems biology approach requires them to learn a common language in order to communicate with each other. This dissertation, performed in a bioinformatics lab at the Department of Electrical Engineering in Leuven (Belgium), would therefore not have been possible without close collaboration with many clinicians, mainly from the University Hospitals in Belgium, the Netherlands, the UK and Sweden.

High-throughput technologies have made it possible to survey the cancer genome with increasing accuracy and resolution. A distinction should be made between causal contributors to the development of cancer pathogenesis and genetic changes that have been accumulated by chance during the cancer's lifespan [20] [57] [368]. Distinguishing causal abnormalities from accumulated changes can only be obtained with an integrative analysis of multiple data types. Because clinicopathological prognostic indicators partly reflect distinct biology and subtypes, the accuracy and power of gene signatures would increase when clinical knowledge is taken into account. However, not only clinical data can enrich predictive signatures. Other new technologies allow the study of genetic and epigenetic changes in DNA, miRNAs and proteins. It is likely that these views provided by various types of genome-wide data contain different, partly

independent and complementary information. A key point in systems biology and a challenge for bioinformatics is therefore the integration of different levels of biological information and an efficient translation of cancer genome information into molecular markers and therapeutics for use in clinical practice. Improvements in technologies thereby add a new dimension to our ability to obtain a more comprehensive and holistic understanding of the complex biology of cancer.

We were inspired by the idea of Lanckriet [197] and others [93] [285] for the integration of multiple types of genomic information, in order to solve a single classification problem with a higher accuracy than would be possible when based on any of the genomic information sources separately. We started with the integration of clinical and microarray data, through which prognosis prediction for breast cancer patients could be improved. Similar conclusions were drawn for lung cancer by others [293], with models including both microarray data and clinical variables outperforming in most cases those without clinical variables. However, due to the availability of much more high-throughput data sources, we extended our integration approach, described in Chapter 3. To our knowledge, this is the first time that a kernel-based integration method has been applied to multiple high dimensional data sets in the patient domain for the study of cancer. Our results will show that using information from different levels in the central dogma improves classification performance and increases predictive power. The data integration approach was first presented at the *International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* in 2007 [79], in 2008 at the *Pacific Symposium on Biocomputing (PSB)* after extension towards multiple high-throughput data sets [78], to result in a full journal publication in *Genome Medicine* [83]. The importance and relevance of the data integration framework follows from its *highly accessed*-status, with more than 3,000 accesses during the first 30 days after publication. An extension with an available software tool for clinical decision support is under review [84].

When more data sources become available in the future, they can easily be added to our integration framework. Some data types, however, have barely been used in a classification setting and require an extra layer of processing before becoming incorporable into our framework. For array CGH data, for example, we present a method in Chapter 5 for the conversion of copy number variations into features usable for classification. The work on the analysis of array CGH data with kernel methods was initially presented at the *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems (KES)* in 2008 [81], followed by presentations at the *Annual Meeting of the Belgian Statistical Society (BSS)* and the *BeNeLux Bioinformatics Conference (BBC)*, both in 2008. A follow-up paper was presented in 2009 at *PSB*, and Chapter 9 in the book *Investigating human cancer with computational intelligence techniques* was dedicated to this work.

Despite mature high-throughput technologies, histopathology will remain a

cornerstone in cancer decision making based on all efforts delivered to date. In Chapter 4 we therefore focus on an improved modeling of clinicopathological parameters by taking into account the type and range of each variable contrary to traditional kernel functions. This work was presented at *EMBC* in 2009 [77] and an extended version is under review as full paper [87].

Although the main focus in this dissertation is on patient data, a large amount of interactome data is freely available, referred to as *secondary data*, with information on, for example, pathway membership, interactions at protein level and miRNAs by which genes are targeted. Such data urgently need to be incorporated in diagnostic and prognostic tools in an intelligent way, in order to improve clinical decision support, to decipher the biological deregulated processes in which key markers are involved and to identify more robust signatures. Because in the past decade the majority of studies only gathered transcriptomics data due to the maturity of microarray technology, in Chapter 6 we have incorporated prior biological knowledge from databases with various, distinct gene information in a classification model based on microarray data. Where similarity between patients is traditionally calculated with a similarity measure based on patients' gene expression profiles, interactome data from any database can be used to improve the way patient similarities are calculated and thus to improve classification performance, irrespective of cancer type and microarray platform. The underlying principle is the exploitation of relations between genes with similar functions but active in alternative pathways. Moreover, considering prior knowledge may benefit investigation of biological functionality. This work was presented in 2009 at *BBC* and a full paper has been published in PLoS ONE [86].

1.6 Chapter-by-chapter Overview

Figure 1.7 gives an overview of how the different chapters are related to each other. An introduction in **Chapter 2** is given on the methodological aspects that are important throughout this dissertation. The main focus is on kernel methods and support vector machines. However, high dimensional data in biomedical applications also necessitate feature selection to increase biological interpretability and decrease cost of clinical decision support. In **Chapter 3**, a kernel-based integration approach for clinical decision support is proposed through which integration occurs within the patient domain before the actual classifier is built. Despite the rise of high-throughput data, for most diseases and examinations clinical data still guide clinical management. **Chapter 4** introduces a new kernel function specifically adapted to the characteristics of such data. Not only clinical but also genomic data need to undergo an extra preprocessing step before becoming incorporable in our kernel-based data integration framework. Because not much attention has been paid to the use of genomic data for classification, in **Chapter 5**

a novel methodological approach is described that combines recurrent hidden Markov Models for the detection of copy number alterations with support vector machines for classification. Finally, our data integration approach can further be improved with non-patient related data obtained from external databases on different aspects of biological systems, such as KEGG and PROSITE. **Chapter 6** gives a description on how relations between genes with similar functions but active in alternative pathways can be exploited based on spectral graph theory, in order to improve clinical decision support. In **Chapter 7**, conclusions and future work are described.

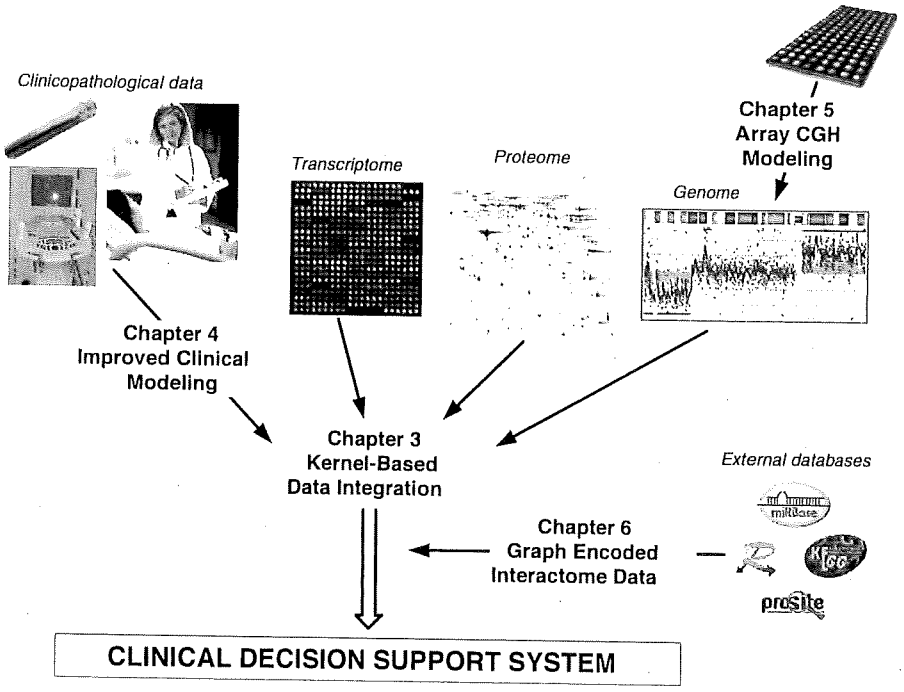


Figure 1.7: *Overview of the relations between the chapters in this dissertation*

1.7 Other Research

The interdisciplinary nature of our research has resulted in a number of research projects that are not explicitly included in this dissertation, but summarized below.

1.7.1 Microarray Studies

Together with Prof. K. Haustermans from the University Hospital in Leuven and Prof. J.-P. Machiels from St Luc University Hospital in Brussels, we investigated proteomics and microarray data from 36 patients with rectal cancer (see Section 3.3) to characterize the molecular pathways activated or inhibited by cetuximab when combined with chemoradiotherapy [78] [83]. To our knowledge, we were the first to investigate more than only microarray data from patients with rectal cancer. Moreover, protein abundance and gene expression were measured at multiple time points during therapy. We identified molecular profiles and potential biomarkers that are currently under investigation for their ability to improve patient selection for cetuximab-based chemoradiotherapy. These results were presented in 2007 at the *International Wolfsberg Meeting on Molecular Radiation Biology/Oncology* and the *International Congress of Radiation Research (ICRR)*, in 2008 at the *American Society of Clinical Oncology Annual Meeting (ASCO)* and have been published in the *Journal of Clinical Oncology* [94].

Lymph node (LN) involvement, which is the most important prognostic factor in breast cancer, was studied in cooperation with Dr. A. Smeets and Prof. M.-R. Christiaens from the University Hospital in Leuven, to enlarge knowledge about the underlying molecular mechanisms. Contrary to breast cancer studies to date and after years of a "one size fits all" approach for patient treatment [189] [314], we considered a very homogeneous group of 96 primary breast tumors (that is, postmenopausal, ER+, HER2-, grade 3 invasive ductal cancer). A molecular signature associated with nodal metastasis was identified. Both an internal and external validation on patients with the same homogeneous characteristics provided evidence that LN involvement in breast cancer is not a complete random process. Our findings were presented at the *IMPACT Breast Cancer Conference* in 2009 and the *European Breast Cancer Conference (EBCC)* in 2010. In cooperation with Dr. D. Lambrechts from the University Hospital in Leuven, we also gathered expression of 229 miRNAs for 83 out of the 96 samples. Eight miRNAs were found to significantly decrease their computationally predicted targets in the LN negative group, while at the same time being inactive in the LN positive group. These miRNAs, hypothesized to suppress LN invasion, were presented at the *IMPACT Breast Cancer Conference* in 2010.

We further assessed the prognostic value of gene expression induced by chronic hypoxia for hepatocellular carcinoma. By combining a liver cell line under chronic hypoxia with independent microarray data, a liver-specific, prognostic 7-gene signature was obtained. This work is the result of cooperation with Prof. J. van Pelt and Prof. C. Verslype from the University Hospital in Leuven and Prof. L. Libbrecht from the University Hospital in Ghent. Our 7-gene signature was presented at the *Belgian Week of Gastroenterology* and the *European Association for the Study of the Liver (EASL)* in 2010, and a full paper is under review

[346]. Prof. L. Libbrecht and Dr. O. Gevaert also participated in a project to evaluate the molecular profile of epithelial human cancer in a large number of samples and tissues, and to investigate whether metastases resemble their tissue of origin or tissue of destination. Our molecular taxonomy of epithelial human cancer indicated correlations among tissues, while metastases from gastrointestinal origin either remembered their tissue of origin or adapted to the tissue of destination. This taxonomy has been published in *BMC Medical Genomics* [138].

The last microarray study that is presented here was performed in cooperation with the Bioinformatics and Neurology group from the Erasmus Medical Center in Rotterdam, and has been published in *Cancer Research* [144]. Because histological classification of glioma is subject to a large inter-observer variability, we applied a hybrid clustering algorithm to microarray data of a large cohort of ~300 glioma samples to obtain a more accurate and objective molecular classification. External validation confirmed that our seven molecular subgroups were different from traditional histological groups with better correlation to patient survival. Our work on hereditary ovarian cancer in cooperation with Dr. K. Leunen and Prof. E. Legius from the University Hospital in Leuven is described in the introduction of Chapter 5.

1.7.2 Proteomic Studies

Within clinical decision support, there is great interest in non-invasive tools (for example, a blood test) to study tumor cells. Proteomic techniques allow detecting changes in serum proteins that are cancer-associated and secreted into the blood. At the University Hospital in Leuven, mass spectrometry was applied by Dr. T. Van Gorp on serum samples of 60 cervical cancer patients to search for tumor-specific serum biomarkers related to LN involvement. Candidate markers were detected and presented in 2009 at the *European Society of Gynecological Oncology (ESGO)*. Early detection of ovarian cancer is also essential due to the discrepancy in survival between early and advanced stage. To date, the most frequently used serum marker for disease monitoring is CA 125. This marker, however, is not approved for detection due to a poor sensitivity of 65%. A combination of CA 125 with human epididymal secretory protein E4 (HE4), on the other hand, distinguishes benign from malignant ovarian masses with a sensitivity of 76%. In collaboration with Dr. T. Van Gorp and Prof. I. Vergote from the University Hospital in Leuven, we performed an independent validation study on preoperative serum levels of CA 125 and HE4 in 389 women with a pelvic mass. Calculation of sensitivity and specificity of both markers and their combination demonstrated similar performance indices as published in specialist literature. These validation results were presented at *ESGO* in 2009 and a full paper has been submitted [340].

Because the study of blood and serum samples is complicated by the presence of

high abundant proteins that mask less abundant and probably more interesting proteins [47], we also investigated tumor tissue. Heterogeneity in tumor tissue requires the use of Laser Microdissection (LMD) [60]. Applying this step prior to any of the discussed technologies allows to obtain a more homogeneous population of tumor cells. Disadvantages are time, the low amount of tissue that remains and loss of information about the stroma and other surrounding tissue. In cooperation with Dr. I. Cadron and Prof. I. Vergote from the University Hospital in Leuven, we showed that with SELDI-TOF clear protein profiles can be obtained from laser microdissected cells, published in *Anticancer Research* [46] and presented in 2008 at the *American Association for Clinical Chemistry Meeting (AACC)* and *Biennial Meeting International Gynecologic Cancer Society (IGCS)*. A similar combination of LMD with SELDI-TOF MS in cooperation with Dr. T. Van Gorp led to novel biomarkers for the prediction of LN involvement in endometrial cancer, presented in 2008 at *IGCS*. In both studies, we investigated peak detection algorithms to convert the spectra to vectors of intensity values, usable as features for classification. In this way, our kernel-based integration framework presented in Chapter 3 becomes applicable to MS data. An example of MS spectra with detected peaks is provided in Figure 1.8.

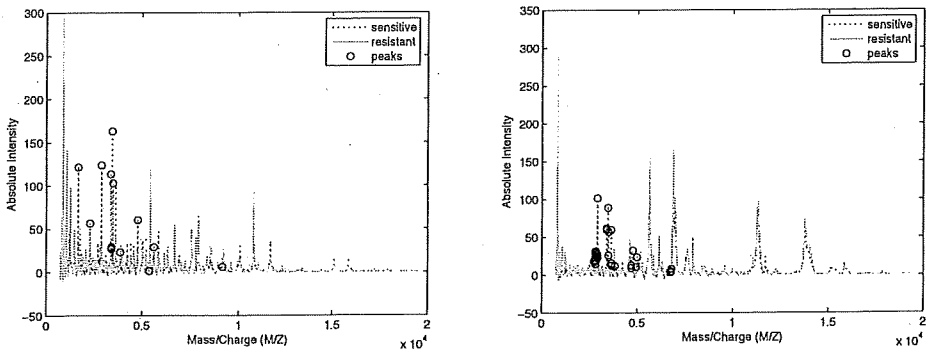


Figure 1.8: *Average spectrum of platinum sensitive (black dotted) and resistant (gray) ovarian cancer samples on chip surfaces (a) CM10 and (b) IMAC30.* (taken from Cadron *et al.*, 2009 [46])

Standard treatment for advanced stage ovarian cancer is surgery combined with platinum-based chemotherapy. 25% of these patients, however, turns out to be platinum resistant and relapses within 6 months after surgery. As no prognostic markers are available yet and understanding of the mechanism of platinum resistance is poor, the sixth framework project OVCAD from the EU aims at identifying biomarkers related to platinum sensitivity and response. In strong collaboration with Dr. I. Cadron, SELDI-TOF MS data from 78 ovarian cancer

patients were studied to investigate the presence of discriminatory protein profiles. The resulting potential biomarkers related to platinum resistance were presented in 2009 at *ESGO*. A full paper will be submitted after internal validation.

1.7.3 Biomedical Studies

We have intensively studied growth in the first trimester of a pregnancy in cooperation with Prof. D. Timmerman and Dr. A. Pexsters from the University Hospital in Leuven, Prof. T. Bourne from Imperial College London, and Dr. C. Bottomley from St George's University of London. Because ultrasound measurements of the crown-rump length (CRL) of an embryo or fetus in the first trimester is more accurate at predicting gestational age than menstrual dating, a growth curve for CRL in function of gestational age was described by Robinson in 1973 [270]. Due to improvements in sonography in the meantime, we constructed a new CRL curve for normal singleton pregnancies. Use of this curve may lead to more accurate estimations of gestational age because comparison with the Robinson curve showed an underestimation of 3 days at the beginning and an overestimation of 2 days at the end of the first trimester when dating a pregnancy based on the Robinson curve. Our data was presented at the *World Congress on Ultrasound in Obstetrics and Gynecology (ISUOG)* in 2008 and 2009, and will appear in *Ultrasound in Obstetrics and Gynecology* [248]. We also compared our CRL curve with data from 134 genetically abnormal singleton pregnancies. At the same conferences, we presented our findings that pregnancies with genetic abnormalities show a smaller than expected CRL, with the most pronounced difference for trisomy 18 pregnancies. A full paper on growth in genetic abnormalities will be submitted [249].

Although fetal growth in normal pregnancies is often assumed to be uniform, we investigated the influence of maternal factors on first trimester growth. Rate of growth increased with advancing maternal age as well as for women of black ethnic origin compared to white. Vaginal bleeding, abdominal pain, obstetric history and body mass index did not influence CRL. These results were presented at *ISUOG* in 2007, 2008 and 2009, and at the *Annual Meeting of the European Society of Human Reproduction and Embryology (ESHRE)* in 2008. The results on maternal age, ethnicity and body mass index have been published in *Human Reproduction* [35] [281].

Reduced growth in the first trimester is associated with an increased risk of miscarriage and adverse late pregnancy outcome. To date, most assessments of early fetal growth have used cross-sectional measurements to compare growth of fetuses at one time point. Several examinations per fetus had to be disregarded to reduce data to one examination per fetus, or else independence was assumed. To overcome the problem of co-dependence, we have introduced the growth assessment

technique Functional Linear Discriminant Analysis (FLDA) for the statistical analyses of longitudinal biomedical data. Traditional LDA classifies data by maximizing the ratio of the between-class variation to the within-class variation. FLDA, originally developed by James and Hastie [173], is an extension of LDA with curves as predictor variables and it can handle irregularly sampled curves and sparse data. This is characteristic for early pregnancy data because the majority of women undergo only 2 or 3 scans at variable times during the first trimester. Serial observations from each individual are modeled with a spline function, represented in Figure 1.9a. Gathering the spline functions per class of women results in a mean coefficient vector, based on which new women can be classified (see Figure 1.9b). We applied this technique to longitudinal CRL measurements to assess and compare growth in pregnancies destined to miscarry with those remaining viable. Due to a significantly different growth rate, FLDA could discriminate normal from abnormal growth for the prediction of miscarriage with high specificity. Importantly, when a pregnancy is followed up with this technique, complications can be detected earlier and prevented where possible. These results were presented at the *Blair Bell Research Society Meeting* in 2007, the *International Meeting on Innovations & Progress in Healthcare for Women* in 2008, at *ISUOG* in 2007, 2008 and 2009, at *ESHRE* in 2008 and have been published in *Human Reproduction* [34].

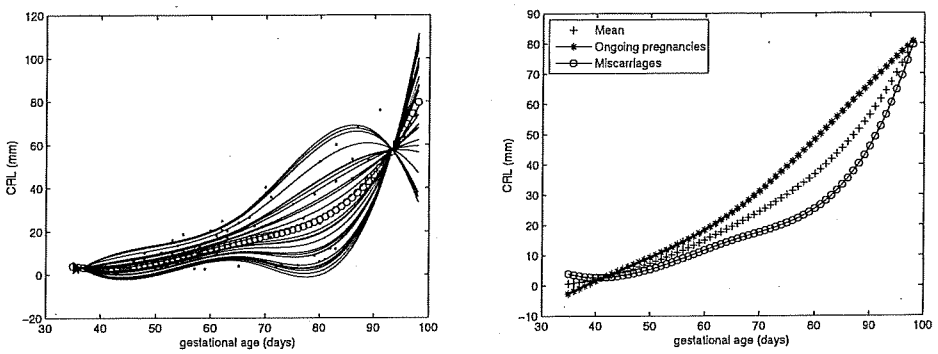


Figure 1.9: small*Illustration of the FLDA technique to discriminate CRL measurements of ongoing pregnancies and miscarriages:* (a) Individual spline curves and centroid (○) for the class of miscarriages; (b) Discriminatory boundary (+) and centroids for the class of ongoing pregnancies (⋆) and miscarriages (○). (taken from Bottomley *et al.*, 2009 [34])

The introduction of the FLDA technique has resulted in a shift in biomedical research. FLDA is not limited to measurements in function of time, but can be applied to decipher the relationship between any two variables. In cooperation

with Prof. C. Lees from Cambridge University Hospitals, Prof. T. Bourne from Imperial College London, and Prof. D. Timmerman from the University Hospital in Leuven, we established the relationship of umbilical artery and middle cerebral artery Doppler impedance with abdominal circumference in babies with severe early onset fetal growth restriction. These relationships, presented at *ISUOG* in 2009 and submitted as full paper [320], may enable a clearer understanding of fetal physiological changes in response to growth restriction. At *ISUOG* in 2009 we also presented our findings on the association of pregnancy associated plasma protein A (PAPP-A) levels in the first trimester with birthweight. First trimester fetal growth rate was not related to birthweight or first trimester PAPP-A levels [150].

Results of other studies in cooperation with Prof. D. Timmerman, Dr. T. Van den Bosch and Dr. C. Van Holsbeke from the University Hospital in Leuven, Prof. T. Bourne from Imperial College London, Dr. E. Kirk from St George's University of London, and Prof. L. Valentin from Malmö University Hospital were presented at *ISUOG* from 2007 to 2009 (see 'Publications by the author' for a full list of abstracts). To date, the majority of these results have been extended to full paper publications in *Ultrasound in Obstetrics and Gynecology* [333] [336] [337] [338] [342] [345], *Journal of Ultrasound in Medicine* [335], *Acta Obstetricia et Gynecologica Scandinavica* [182], *Gynecologic and Obstetric Investigation* [341], and *Facts, Views & Opinions in ObGyn* [334]. Results from our collaboration with the International Ovarian Tumor Analysis Group are described earlier in Section 1.3.

Chapter 2

A Methodological Primer

A distinction can be made between supervised and unsupervised learning, depending on whether scientific or clinical characteristics of the training samples are available and used. In cancer research, diagnosis, prognosis and decisions on treatment can be improved by learning a relation between pattern x^i (also referred to as sample, instance or observation) and the corresponding label, target or output y_i. This task is an instance of supervised learning, known as classification. The extracted knowledge can subsequently be applied to new patterns x for accelerating the availability of y before surgery or for early cancer detection, thereby improving survival and quality of life. Regression can be considered as an extension towards estimating real-valued functions f for a continuous outcome vector y, in such a way that $y = f(x)$. For binary classification, the prediction model follows from applying the sign-rule, that is $y = sign[f(x)]$. In this dissertation, kernel methods and support vector machines are used for supervised learning. Due to the typical high-throughput data sets in biomedical applications, selecting the most relevant features for incorporation in another classifier is a necessity. This will additionally decrease cost of clinical decision support.

2.1 Evaluation Measures

In machine learning and statistics, two classes of samples that need to be distinguished are often referred to as the positive and negative class. The meaning of positive and negative depends on the specific application area. As our focus is

on cancer research, the positive class is often annotated with disease samples or samples with a malignant tumor, while the negative class contains normal samples or samples with a benign tumor. In studies in which two cancer subtypes are investigated, the allocation of the subtypes to the positive or negative class is arbitrary.

For the evaluation of a binary classifier, the predicted labels are compared with the given output labels, thereby distinguishing correctly and incorrectly classified samples from both classes. These categories are indicated as true positives (TP) (that is, the number of correctly classified positive samples), false positives (FP) (that is, the number of negative samples that are incorrectly classified as positive), true negatives (TN), and false negatives (FN). Several evaluation measures shown in (2.1) can be derived from these figures [26]. The accuracy represents the proportion of N samples that are correctly classified. The sensitivity or true positive rate equals the proportion of positive samples that are correctly classified as such, while specificity, also known as the true negative rate, measures the proportion of negative samples identified as negative. As it is often desirable for a test to have both high sensitivity and specificity, the Youden's index J, proportional to the sum of sensitivity and specificity can be maximized. For a perfect test, J equals 1, while the minimal value is 0 for a test lacking diagnostic value. Another measure that combines sensitivity and specificity is the odds ratio (OR), defined as the ratio of the positive likelihood ratio (LR+) to the negative one (LR−). LR+ and LR− represent the increase in odds favoring a positive outcome when the classifier assigns samples to the positive and negative class, respectively. A high LR+ and low LR− value are indicative for a useful test, corresponding to an OR above 1 so that the positive outcome is more likely to occur in case of a positive test result and vice versa for the negative outcome. The balanced error rate (BER) is the average of the misclassification rate per class. BER coincides with the classification error (that is, 1-accuracy) when both classes contain the same number of samples. In case of unbalanced classes, BER assigns a different weight, inversely related to the size of each class, to false negative and false positive errors. A majority vote model corresponds to a BER equal to 0.5 (for example, TP=N_P, FP=N_N, FN=0, TN=0 with N_P the number of samples in the positive class, N_N the number of samples in the negative class and $N_P \gg N_N$).

$$\begin{aligned} Accuracy &= (TP+TN)/N \\ Sensitivity &= TP/(TP+FN) \\ Specificity &= TN/(TN+FP) \\ J &= Sensitivity + Specificity - 1 \end{aligned} \tag{2.1}$$

$$OR = \frac{LR+}{LR-} = \frac{sensitivity}{1-specificity} / \frac{1-sensitivity}{specificity}$$

$$BER = 0.5 \times \left(FP/(TN+FP) + FN/(TP+FN)\right)$$

For all previous performance criteria, samples are assigned to the positive or negative class by binarizing the continuous prediction of a classifier. These criteria do therefore not provide a unique description of diagnostic performance. They depend on the discrimination threshold, arbitrarily chosen or based on disease-specific requirements (for example, high sensitivity or high specificity). The receiver operating characteristics (ROC) curve [26] [226] summarizes the performance of a classifier by showing the true positive rate (sensitivity) versus false positive rate (1-specificity) as the discrimination threshold is varied. Each threshold corresponds to a particular operating point on the ROC curve. Increasing the threshold simultaneously enlarges the number of true positives and false alarms, whilst reducing the number of false positives automatically reduces the number of hits. The area under the ROC curve (AUC) [154] is an evaluation measure that is independent of the operating point. An AUC of 0.5 corresponds to a random classifier lacking discriminative power, while a perfect classifier is characterized by an AUC of 1. Thus, the larger the AUC, the better the classifier with a higher sensitivity at a small false positive rate (high specificity). In this dissertation, the AUC will be used as a performance criterion and an evaluation measure. Comparing AUC values in different classifiers is less sensitive to the specific threshold used for assigning observations to a class.

Although our main focus is on classification, the toolbox for the integration of data sets provided in Chapter 3 can also be used for regression problems. For those problems, the mean squared error (MSE) was considered and defined as:

$$MSE = (1/N) \times \|y - \hat{y}\|_2^2, \tag{2.2}$$

with y the true, known output, $\hat{y}$ the predicted outcome vector, and $\| \ \|_2^2$ denoting the squared Euclidean norm.

2.2 Classification Methods

After having introduced some evaluation measures, a bottom-up overview is provided on classification methods. A general introduction on kernel methods is followed by an explanation on kernel algorithms for supervised classification.

2.2.1 Kernel Methods

Kernel methods are a group of algorithms that can handle a very wide range of data types, such as vectors, sequences and networks [285]. As shown in Figure 2.1, they work by embedding the data x into a high dimensional feature space with the nonlinear feature map $\Phi(x)$. However, an explicit construction of Φ and computations in the feature space can be evaded by applying the *kernel trick*, that is $k(x^i, x^j) = \langle \Phi(x^i), \Phi(x^j) \rangle$. Data are not represented individually but through a set of pairwise comparisons in the kernel matrix, obtained via the mathematical object $k(x^i, x^j)$. This *kernel function* efficiently computes the inner product $\langle \Phi(x^i), \Phi(x^j) \rangle = \Phi(x^i)^T \Phi(x^j)$ between all pairs of data items x^i and x^j in the feature space. The kernel matrix thus fuses information about the data and the applied kernel function, thereby containing all necessary information for the learning algorithm. The size of the kernel matrix is only determined by N (that is, the number of data items), independent of the nature or complexity of these items. For example, a set of 100 patients each characterized by 6,913 gene expression values is represented by a 100 $\times$ 100 kernel matrix; similarly characterization by 96 proteins with a 3D structure results in a 100 $\times$ 100 kernel matrix. The representation of any data set by this real-valued square matrix makes kernel methods ideally positioned for heterogeneous data integration.

If the kernel function, and thus the feature space, is chosen suitably, the embedded data are linearly separable and any learning algorithm can easily be applied in the embedded space. On the other hand, if the data are embedded in a space with too many irrelevant features, the mapped data points will be orthogonal to each other, the kernel matrix becomes diagonal and all structure is lost [71]. Any symmetric, positive semidefinite function is a valid kernel function, resulting in many possible kernels. However, no formal proof of optimality exists for the use of one kernel function above an other. The functions that are most frequently employed in classification problems are the linear kernel ${x^i}^T x^j$, the polynomial kernel $({x^i}^T x^j + \tau)^d$ with - as kernel parameters - the intercept constant $\tau \in \mathbb{R}^+$ and degree $d \in \mathbb{N}$, and the radial basis function (RBF) $\exp\big(-\|x^i - x^j\|_2^2/\sigma^2\big)$ with $\sigma \in \mathbb{R}^+$ representing the width of a Gaussian distribution centered on the data points. The polynomial kernel corresponds to a feature space spanned by all products of d variables at the most. This kernel results in a quadratic separating surface in the input space for $d = 2$, and it represents the cubic kernel for $d = 3$. These kernel functions all correspond to a different transformation of the data, meaning that they extract a specific type of information from the data.

For each kernel function, a normalized version can be defined as

$$\tilde{k}(x^i, x^j) = k(x^i, x^j) / \sqrt{k(x^i, x^i) k(x^j, x^j)}. \tag{2.3}$$

With this normalized version, the data points are projected onto the unit sphere

in such a manner that the elements in the kernel matrix are bounded. Moreover, as all diagonal elements in a normalized kernel matrix equal 1, a normalized kernel can be interpreted as a measure of similarity, a concept that will be important in Chapter 4. For any two objects x^i and x^j, a kernel function can be written as:

$$k(x^i, x^j) = \frac{\|\Phi(x^i)\|^2 + \|\Phi(x^j)\|^2 - d(\Phi(x^i), \Phi(x^j))^2}{2} = \langle \Phi(x^i), \Phi(x^j) \rangle, \quad (2.4)$$

with $d(u,v)^2 = \langle (u-v), (u-v) \rangle$ and $\|u\|^2 = \langle u, u \rangle$. The kernel $k(x^i, x^j)$ thus measures the similarity between x^i and x^j as the opposite of the square distance between their representations in the feature space, up to the terms $\|\Phi(x^i)\|^2$ and $\|\Phi(x^j)\|^2$. A normalized kernel lacks the gap between the notion of inner product and similarity because all data points have the same length in the feature space (that is, $k(x,x) = \|\Phi(x)\|^2$ is constant for all x).

2.2.2 Support Vector Machines

We will now introduce a kernel algorithm for supervised classification. Consider a training set for classification $\{x^i, y_i\}_{i=1}^N$ of N samples with feature vectors $x^i \in \mathbb{R}^p$ and binary output labels $y_i \in \{-1, +1\}$. The aim is to train a function $f(x) = y$ that correctly classifies unseen samples $\{x, y\}$. Data points x^i with $f(x^i) \geq 0$ are assigned the label +1, data points with $f(x^i) < 0$ the label -1. Imposing no restriction on the set of functions from which the estimated f is chosen can result in selecting a function with good training performance but bad generalization to unseen data. The set of functions thus needs to be restricted to functions with a capacity that suits the amount of available training data. At the same time, the set should be large enough to provide functions that can model the dependencies in the data. Vapnik [348] considered as set the class of hyperplanes $\{\langle w, x \rangle + b = 0\}$, corresponding to a linear function of the form $f(x) = w^T x + b$. Variable b represents the bias term and the normal vector w of the decision hyperplane corresponds to weights assigned to the p variables. Vector w can also be represented as a general linear combination of the training data points. Vapnik further introduced a principled way to choose the best possible hyperplane among all hyperplanes that separate two classes of samples. In the simplest case of linear separable data, the corresponding classifier was defined as $y(x) = \text{sign}[w^T x + b]$, with the closest data points x^i satisfying the equality constraint $|w^T x^i + b| = 1$. This optimal discriminant boundary is known as the linear Support Vector Machine (SVM). The above constraint forces the margin (that is, the distance between the nearest points of both classes) to be equal to $2/\|w\|_2$. The SVM classifier is thus obtained from the solution to a convex optimization problem in which the margin is maximized or equivalently $\|w\|_2 = w^T w$ minimized, subject to the constraint of correctly classifying the training data points with strong confidence.

The optimization problem in primal weight space, which is linear in the parameters w and b and characterized by a unique and global solution, is given by:

$$\min_{w,b} \left(\frac{1}{2} w^T w\right) \qquad \text{s.t.} \quad y_i[w^T x^i + b] \geq 1, \; i = 1, \ldots, N \tag{2.5}$$

As the dimension of w is determined by the number of variables, the calculation of w can be avoided for problems with high dimensional data by solving the problem as a quadratic programming (QP) problem in dual space. The unknown variables in dual space are the Lagrange multipliers α_i, referred to as *support values*. Vector α is of size N and thus typically small for high-throughput data sets. The resulting classifier rule equals $y(x) = \text{sign}[\sum_{i=1}^{N} \alpha_i y_i x^{i^T} x + b]$. A nice property of the SVM is sparseness with many of the support values forced to zero. Only the training data points with non-zero α_i, referred to as *support vectors* and located close to the decision boundary, are required for the classification of future data. The decision boundary can thus be expressed in terms of the limited number of support vectors, thereby removing the influence of data points that are far away from the decision boundary.

In most binary classification problems, classes are characterized by overlapping distributions in such a way that a separating hyperplane does not exist. Even when the data are separable, the hyperplane obtained with problem formulation (2.5) may not be the best solution in presence of outliers or mislabeled data points. To obtain an algorithm that can cope with misclassifications and a certain fraction of outliers, problem (2.5) is extended with slack variables ξ. In this way, the focus is more on the confidence of classification than on the number of misclassifications. For linear non-separable data, the decision boundary can directly be constructed in the input space and the problem formulation in primal space changes into:

$$\min_{w,b} \left(\frac{1}{2} w^T w + c \sum_{i=1}^{N} \xi_i\right) \tag{2.6}$$

$$\text{s.t.} \quad y_i[w^T x^i + b] \geq 1 - \xi_i, \; i = 1, \ldots, N$$

$$\xi_i \geq 0, \; i = 1, \ldots, N$$

The positive hyperparameter c represents the trade-off between the requirement of a large margin and that of few misclassifications (that is, $\xi_i > 1$) or classifications with little confidence ($0 < \xi_i \leq 1$). Increasing c forces the function $f(x) = w^T x + b$ to have a smaller margin, but correctly classifying more data points with strong confidence.

In case of nonlinear separable or non-separable data, a transformation of the data from the input space into a high dimensional feature space is required. In the latter space, the classes are assumed to be linearly separable in case the nonlinear feature map Φ is properly chosen. The constructed hyperplane corresponds to the nonlinear discriminant boundary $y(x) = \text{sign}[w^T\Phi(x) + b]$ in the original input space. This principle is represented in Figure 2.1. The linear SVM is thus easily extendable to the nonlinear case by replacing x^i by $\Phi(x^i)$ in problem formulation (2.6) and applying the kernel trick where possible to avoid explicit knowledge of $\Phi(x)$ and explicit computation in the feature space. However, as $\Phi(x)$ can be infinitely dimensional, the same applies to w. The nonlinear SVM case therefore needs to be solved in dual space with as discriminant boundary $y(x) = \text{sign}[\sum_{i=1}^{N} \alpha_i y_i k(x^i, x) + b]$. When the RBF kernel function is used, the discriminant boundary corresponds to a sum of Gaussian distributions, centered around the support vectors. Large σ values result in smooth Gaussians, and therefore a smooth or almost linear decision boundary. When σ is decreased, the Gaussians become more peaked around the support vectors, leading to more complex decision boundaries. When the width is too small, the data are practically memorized, resulting in overfitting. In analogy to the linear (non-) separable case, support vectors lie on the margin with $\alpha_i = c$ for misclassifications, or on the ± 1 lines around the hyperplane when $0 < \alpha_i < c$.

Support vector machines are also suited for regression problems. In case of a linear problem, the function $f(x) = w^T x + b$ is estimated by using the training data with continuous output values $y \in \mathbb{R}$. Similarly to classification, this strategy can be extended to nonlinear regression problems by replacing x by $\Phi(x)$. The resulting SVM model becomes $y(x) = \sum_{i=1}^{N} \alpha_i k(x^i, x) + b$. As the focus in this dissertation is on classification problems, we refer to [318] for a more detailed description of SVMs for function estimation.

2.2.3 Least Squares Support Vector Machines

A modified version of SVM, the Least Squares Support Vector Machine (LS-SVM), was developed by Suykens *et al.* [318] [319]. The main difference between the standard SVM and the LS-SVM is that the inequality constraints in the formulation of the SVM (2.6) are replaced by equality constraints, thereby considering the value 1 at the right hand side as a target value rather than a threshold value. In this way, the interpretation of the slack variables ξ_i slightly changes into the offset with respect to target 1, and the variables are referred to as error variables e_i. Moreover, the second term in the objective function is replaced by the squared error contribution. These modifications transform the QP problem, which requires expensive computation, into a much simpler set of linear equations. The LS-SVM is therefore much faster on high dimensional data sets

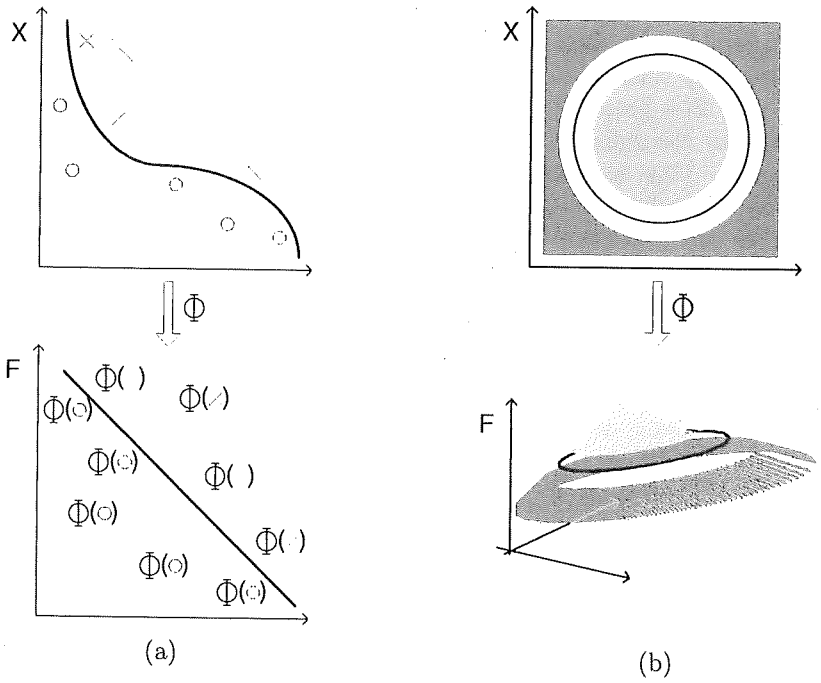


Figure 2.1: *Principle of the SVM for nonlinear data (○ and ×): X* represents the original input space, F a high dimensional feature space determined by the kernel function or choice of Φ. After a transformation from X to F, a linearly assumed classification problem is solved, resulting in a linear hyperplane in F and a nonlinear discriminant boundary in X (thick line).

with low computational cost. The formula for this constrained convex optimization problem equals:

$$\min_{w,b,e} \left(\frac{1}{2} w^T w + \gamma \frac{1}{2} \sum_{i=1}^{N} e_i^2 \right) \tag{2.7}$$

$$\text{s.t.} \qquad y_i [w^T \Phi(x^i) + b] = 1 - e_i, \; i = 1, \ldots, N$$

with γ the regularization parameter, representing - in analogy to hyperparameter c in (2.6) - the trade-off between margin maximization and minimization of the squared error contribution. The latter can be interpreted as the minimization of the within-class scatter for the two classes with target values +1 and -1. When the function $f(x) = w^T x + b$ fits the training data well, this term is small. The regularization term $w^T w$, on the other hand, ensures that the solution of problem

(2.7) is smooth, or in other words that the obtained function f slowly varies between similar data points with respect to the geometry defined by the kernel function. In case of noisy data, a powerful protection against overfitting is forcing smoothness and a large margin on the training data and allowing larger within-class scatter by keeping γ small. This also ensures good test performance. A large γ, on the other hand, will lead to a better fit of the model to the training data, thereby reducing the error term and within-class scatter but increasing complexity. In conclusion: regularization makes classification robust to uncertainties in the data and has been shown to be very important for classification of high dimensional data to ensure good generalization, even with linear classifiers [254].

In dual space, the equivalent problem is obtained by solving the Lagrangian, resulting in a system of linear equations in function of the number of data points N:

$$\left[\begin{array}{c|c} 0 & y^T \\ \hline y & \Omega + I/\gamma \end{array}\right] \left[\begin{array}{c} b \\ \hline \alpha \end{array}\right] = \left[\begin{array}{c} 0 \\ \hline 1_N \end{array}\right], \tag{2.8}$$

with α and b the unknown variables and $\Omega_{ij} = y_i y_j k(x^i, x^j)$, $i, j = 1, \ldots, N$. Because for high dimensional data sets N is typically low, all experiments in this dissertation were performed in dual space.

The benchmarking study of Van Gestel and colleagues [339] on 20 UCI data sets has revealed that the LS-SVM is capable of achieving a performance that is comparable to the standard SVM. LS-SVM classifiers have also achieved satisfactory classification accuracies on microarray data [254]. The improved implementation efficiency of the LS-SVM, however, is at the expense of sparseness in the solution vector. As the support values are proportional to the error variables (that is, $\alpha_i = \gamma e_i$), all training data points are support vectors and contribute to the model. The points with largest $|\alpha_i|$ are located close and far from the decision boundary. The sparsity in data points is thus lost with respect to the SVM. However, pruning methods can be applied to the sorted support value spectrum, thereby removing less meaningful data points and recomputing the LS-SVM on the reduced set of training points to obtain a sparse representation of the original LS-SVM system [317]. Similar as with SVMs, the LS-SVM formulation is applicable to regression problems [318].

2.2.4 Weighted LS-SVM

In many two-class problems, data sets are skewed in favor of one class with $N_P \gg N_N$ or $N_N \gg N_P$. The contribution of false negative and false positive errors to the performance assessment criterion is therefore not balanced. We used a weighted LS-SVM (wLS-SVM) in which a different weight ζ_i is given to positive and negative

samples in order to take this unbalance into account [51]. The objective function of problem (2.7) changes into:

$$\min_{w,b,e} \left(\frac{1}{2} w^T w + \gamma \frac{1}{2} \sum_{i=1}^{N} \zeta_i e_i^2 \right) \tag{2.9}$$

with

$$\zeta_i = \begin{cases} \frac{N}{2N_P} & \text{if } y_i = +1 \\ \frac{N}{2N_N} & \text{if } y_i = -1 \end{cases}$$

The choice of ζ is asymptotically equivalent to a re-sampling of the data till an equal number of positive and negative samples is obtained. In dual space, the introduction of the weighting scheme corresponds to replacing I in (2.8) by $\text{diag}(\zeta_1^{-1}, \ldots, \zeta_N^{-1})$. When interested in explicitly assigning samples to one of both classes for an unbalanced problem in clinical decision support, a bias term correction [318] should be applied, after which $f(x) = 0$ is the operating point. Problem (2.9) should not be confused with the weighted LS-SVM proposed by Suykens and colleagues [317]. Contrary to the SVM, linearity of the objective function in the slack or error variables is lost with the LS-SVM, leading to a weakened outlier resistance. The aim in [317] was therefore to obtain robust estimates in case of outliers or when the underlying assumption of a Gaussian distribution for the error variables e_i is not realistic.

2.3 Feature Selection

Feature selection is required before building the actual models because the prediction accuracy might improve by excluding redundant and non-discriminatory features in high-throughput data sets. Moreover, a limited number of genes and molecules are involved in cancer and knowledge about more relevant candidate features can improve insight into the underlying processes that generated the data. To deal with these problems of large dimensionality and small sample size, a wealth of feature selection techniques has been developed in the domain of machine learning. The choice of the feature selection technique, however, remains a widely discussed topic as no consensus technique exists with an optimal performance on all data sets [277].

A distinction is made between univariate and multivariate feature selection methods. Univariate approaches are computationally simple and fast for high dimensional data, but ignore feature dependencies and are unable to remove redundant features. Multivariate selection techniques, on the other hand,

take correlation information among features into account, but are often not appropriate for high dimensional data sets to discover these underlying complex multivariate correlations due to sample size limitations. Overly-complex feature selection algorithms specifically developed to extract these structures are prone to overfitting. In this dissertation, we opted for univariate methods because univariate feature rankings are easy to interpret, it has been shown that these methods lead to good and stable performances across many cancer types, and in many cases they even yield consistently better results than multivariate approaches [194].

2.3.1 Univariate Feature Selection

An enumeration of univariate feature selection approaches is given in [194]. Despite the wealth of available methods, biologists prefer two of the earliest approaches, being fold changes and t-statistics, due to their simplicity and interpretability. According to Witten and colleagues [365], the question which of both methods to use is of biological nature and should depend on the interest in an absolute change in features or in a change relative to the underlying noise in the data. The biological question thus determines the type of difference between two classes of samples that has biological relevance. From a statistical point of view, however, there is no reason to choose one univariate method over an other. The best performance technique depends on the data set used, with the choice of method dramatically affecting the identified set of features. Moreover, no comparisons have been made across a sufficiently wide range of benchmark data sets. Also Dudoit and colleagues [101] observed that differentially expressed genes vary more between methods than within methods for different significance thresholds.

2.3.2 Differential Expression via Distance Synthesis

To generalize feature selection to any biological system of interest, to get rid of some of the arbitrariness according to which univariate methods are chosen for high dimensional data, and to enhance the robustness and stability of the final feature subset, we used the ensemble feature selection approach DEDS (Differential Expression via Distance Synthesis) [373]. This technique is based on the integration of different test statistics via a distance synthesis scheme. The underlying principle behind DEDS is that features which are highly ranked by multiple statistics are more likely to be expressed differentially than features highly ranked by a single test statistic, thereby reducing the number of false positive findings. The Bioconductor package DEDS in R was used for the combination of the following statistical tests: ordinary fold changes, ordinary t-statistics, SAM statistics and moderated t-statistics. The performance of DEDS in feature ranking is comparable with the best individual statistic, which is in practice often unknown

and depends on the data set used. Additionally, DEDS is not adversely affected by the worst performing statistic and achieves robustness properties that are lacked by the individual statistics [373]. It was therefore used for all high-throughput data sets throughout this dissertation. From our experience, however, DEDS was less appropriate for the low-dimensional proteomic data set used in Chapter 3. As for this data set with a limited number of features one test statistic sufficed, the Wilcoxon rank sum test was opted for.

2.3.3 Optimization of the Feature Subset

As the class-conditional densities are not completely known and only a finite number of samples is available, the number of features cannot be increased arbitrarily and performance starts deteriorating above a certain number of features, known as the *peaking phenomenon* [53]. It has been shown by Sima and colleagues [301] that the peaking phenomenon not only depends on feature characteristics and sample size, but also on classifier type. For SVMs, for example, performance barely or slowly increased with increasing number of features. Moreover, the expected percentage of noisy features in high dimensional data is large. The number of informative features will therefore be dominated by noisy features when performing feature selection [301]. In this dissertation, the set of included features was iteratively enlarged according to the obtained feature ranking, without including more features than the number of samples on which the optimal number of features was determined [206]. As indicated in Chapter 1, the aim of feature selection is not limited to obtaining good performing features, but it additionally requires features to be of biological relevance. As the selected features can highly depend on the specific set of training samples, feature selection was repeated on different subsets for training using LOO or v-fold CV. The features that were most frequently selected in the different iterations are likely to be more relevant candidate features.

2.4 Parameter Setting Methods

We will end this chapter with an overview of methods used to obtain classification accuracy and to estimate the parameters of a classification model [172].

2.4.1 Error Estimation Methods

Using all samples for both model building and model selection results in an optimistically biased estimate and should be avoided. Ideally, when the number of samples is sufficiently large, data can be divided into a training, validation

and test set. The validation set is then used to determine the parameter set at which optimal validation performance is obtained. This limits the modeling of noise in the training data and reduces the influence of outliers on model building. The final model is tested on the independent samples in the test data. In most situations, however, the number of samples is limited and no samples can be fully left untouched. In that case, some form of resampling is required. A distinction is made between holdout, v-fold cross-validation (CV) and leave-one-out (LOO). In the *holdout* method, data are split into a training and test set. Stratification to outcome ensures that the relative proportion of outcomes sampled in both training and test set is similar to the original proportion in the full data set. Models are built on the training set and their performance is assessed on the test set. A critical point, however, is the partitioning. When the training set is small, the resulting classifier is not very robust with a low generalization ability. A small test set, on the other hand, lowers the confidence in the obtained performance. Moreover, this method leads to a pessimistically biased estimate with different error estimates for different random splits. Due to dependence of test performance on the choice of training and test set and to obtain a better estimate of the prediction performance, the split of the data should be repeated a sufficient number of times, referred to as *randomization*. The optimal model is the one that in general performs best for all test sets.

Due to the overlap between test sets in different runs, *v-fold CV* is another possibility. The collection of N samples is split into v folds, possibly stratified to outcome. The samples in $v-1$ folds are used for the construction of the classification model, followed by validation of this model on the left out fold containing N/v samples. This procedure is repeated v times such that each fold, and thus each sample, is once used for testing. Classifier performance is assessed as the overall performance obtained on all v folds. The *leave-one-out* method is a special instance of v-fold CV with v equal to the number of samples. With LOO, one sample is left out for testing in each iteration and the procedure of training and testing is repeated N times. Both v-fold CV and LOO are characterized by high variance on the estimates, but the bias with LOO is smaller as all samples except one can be used in each iteration for training. The specific choice of the estimation method, however, depends on the sample size.

2.4.2 Grid Search

Applying SVMs or LS-SVMs in a fully automatic manner is challenging because questions regarding parameter selection and choice of the kernel function remain largely empirical in real-world applications. As a good generalization performance can be achieved by carefully tuning the algorithm and its parameters [161], an exhaustive grid search was performed in this dissertation for parameter optimization. With r the number of parameters to be optimized, an r-dimensional

grid of different values for those parameters was defined. For the selection of regularization parameter γ, 40 possible values were considered ranging from 10^{-4} to 10^6 on a logarithmic scale. In case of a nonlinear kernel function, the grid was expanded with the additional tuning parameters τ, d and σ. We made the kernel parameters τ and d of the polynomial kernel vary on a linear scale. Width σ of the RBF kernel causes a trade-off between data fit and model complexity, with larger values implying a smoother decision surface. As $\|x^i - x^j\|_2^2$ in the RBF kernel is proportional to the number of variables, scaling σ with the square root of p leads to the incorporation of less nonlinearities into the model in case of high dimensional data, contributing to a reduced risk of overfitting. We therefore took into account the dimensionality of the data for the range of σ, defined on an exponentially growing scale as suggested by Hsu and colleagues [161]. Besides these kernel parameters, the number of selected features and the weights assigned to different data sets are other potential parameters that need to be optimized. With LOO or v-fold CV, in each iteration feature selection was performed where necessary and models were built for all possible combinations of parameters on the defined r-dimensional grid. The optimal set of parameters was chosen corresponding to the model with the highest LOO or v-fold CV AUC. If multiple models had the same maximal AUC, the model with the lowest BER and maximal Youden's index was chosen. In case independent test data were available, the LS-SVM classifier was reconstructed on the complete training set by using the optimal parameters before testing.

2.5 Conclusion

In this chapter, the important concept of the ROC curve was introduced, enabling the evaluation of classifiers independent of the specific application area. Classification methods were subsequently presented in a bottom-up overview. Starting with a general introduction on kernel methods, we evolved from SVMs to LS-SVMs, to finally present a special version of the LS-SVM that takes into account the unbalance, common in biomedical applications. As feature selection technique, we opted for the integration of a collection of univariate feature selection statistics via a distance synthesis scheme. Finally, the grid search approach was chosen for parameter optimization to achieve good generalization performance. Throughout chapters 3 to 6, it will become clear that our justified choices for the AUC measure, wLS-SVM classifier, feature selection method DEDS and the grid search approach can contribute to improved clinical decision support.

Chapter 3

High-Dimensional Kernel-Based Data Integration

The transcriptome does not completely reflect the underlying biology due to alternative splicing, post-translational modifications and the influence of pathological conditions on transcription and translation. The current increase in the amount of available omics data enables integration of many genome-wide data sources such as the genome, transcriptome and proteome. We propose a kernel-based integration approach for clinical decision support through which integration occurs within the patient domain before the actual classifier is built. This approach was followed for two cancer cases, one on rectal cancer with transcriptomics and proteomics data, and one on prostate cancer with transcriptomics and genomics data. For both cancer sites, prognosis and therapy response prediction improved when more than one genome-wide data set was considered. This suggests that integrating multiple genome-wide data sources increases the predictive performance of clinical decision support models. The presented results emphasize the need for comprehensive multi-modal data, to ultimately obtain cost-efficient models from which patients can directly benefit.

3.1 Introduction

Within the field of clinical decision support, there is a huge need for simultaneously analyzing multiple data sets, obtained from the same patients and containing partly independent and complementary information about different aspects of

biological regulation. For a decade, microarray technology has been used in research to improve decision support. Especially within the domain of cancer, it has earned a prominent place with several gene signatures currently being validated in clinical trials. The superiority of gene signatures over clinicopathological parameters, however, has not yet been completely established [107]. Most microarray-based prognostic studies have prompted too optimistic results with much poorer performances on truly independent data. Predictions obtained with molecular signatures applied to the same data set are quite discordant and have been shown to be very unstable [189]. Analyses at multiple hierarchical levels of biological regulation are therefore a necessity and have become possible due to the expansion of available *-omics* data from transcriptomics and genomics to proteomics, metabolomics and epigenomics. In cancer research, however, the number of available data sets is typically limited due to cost.

Despite availability of these technologies, joint analysis is at a preliminary stage. We therefore investigated how multiple high-throughput data sets gathered from the same patients can be combined, in order to exploit the strength of each data set and to efficiently improve clinical decision support [78] [79] [83]. Our starting point was to fuse distinct experimental data sets into one mathematical model for the development of more homogeneous classifiers. This has resulted in a kernel-based integration approach in which fusion occurs within the patient domain. Instead of merging data sets or analyzing them in turn, we thus opted for the actual integration of measurements in multiple data sets. Variables from different omics data sets were treated equally so as to select the most relevant features from each data set.

Although there already have been some studies on data integration (see Section 3.2), our definition for the combination of heterogeneous biological data differs from what has been described so far in specialist literature. In Section 3.4, we will present an approach in which high-throughput data sets are integrated over multiple layers in the genome by treating variables from the different data sets equally. The HIgh-DImensional Data Integration Toolbox HI-DID-IT has been made available and is described in Section 3.5. Two cases summarized in Section 3.3 were considered to test our strategy. In the first case on rectal cancer, tumor regression grade, lymph node status and circumferential margin involvement (CRM) were predicted for 36 patients based on transcriptomics and proteomics data, gathered at two time points during therapy. The second case on prostate cancer involves transcriptomics and genomics data of 55 patients. Tumor grade, stage, metastasis and occurrence of recurrence were available for prediction [198] [199]. In Section 3.6 we would like to demonstrate that appropriately integrating multiple data sets in the patient domain increases predictive power compared to models built on only one data set. Finally, conclusions are given in Section 3.7.

3.2 Review on Data Integration

Other research groups have already proposed the idea of data integration, but most groups only investigated the integration of clinical and microarray data. Tibshirani and colleagues [322] proposed such a framework by reducing microarray data to one variable, addable to models based on clinical characteristics such as age, grade and tumor size. Nevins and colleagues [235] combined clinical risk factors with metagenes (that is, the weighted average expression of a group of genes) in a tree-based classification system. Wang *et al.* [358] combined microarray data with knowledge on two clinicopathological variables. The development of gene signatures was restricted to patients for whom the clinicopathological variables were not sufficient to predict outcome.

A further evolution could be seen in studies in which two -omics data sets were simultaneously considered, in most cases microarray data combined with proteomics or array CGH data. Extensive literature on such studies involving data integration already exists. However, the definition for the integration of high-throughput data sets as it is used in specialist literature differs from our point of view.

In a first group of integration studies, heterogeneous data from different sources were analyzed sequentially [223]; that is, one data set was analyzed while the second was used to confirm the obtained results, to further deepen the understanding of the results [135], to overcome the incomplete coverage and low sensitivity or specificity of the individual experimental approaches [232], or to significantly reduce the number of candidate markers in follow-up studies. Such approaches were used for biological discovery and for a better understanding of the development of a disease, rather than for predictive purposes. For example, Fridlyand and colleagues [124] found three breast tumor subtypes with a distinct CNV pattern based on array CGH data. Microarray data were subsequently analyzed to identify the functional categories that characterized these subtypes. Tomioka *et al.* [326] analyzed in a similar way microarray and array CGH data of patients with neuroblastoma. Genomic signatures resulted from the array CGH data, while molecular signatures were found with microarray analysis. The authors suggested that a combination of these independent prognostic indicators would be clinically useful.

Data integration methods, on the other hand, have been proposed in the database literature as *integration architectures* (see [225] for an overview of these architectures). The term data integration is thus also a synonym for data merging through which different data sets are concatenated at the database level by cross-referencing the sequence identifiers. Data merging requires semantic compatibility among data sets [142] [360], and is a complex task due to, for example, the use of different identifiers, the absence of a 'one gene - one protein' relationship,

alternative splicing, and measurement of multiple signals per gene. In most studies, the concordance between the merged data sets and their interpretation in the context of biological pathways and regulatory mechanisms were investigated. Analyses of the merged data set by clustering or correlating the protein and microarray data can help to identify candidate targets when changes in expression occur at both the gene and protein level. However, there has only been modest success from such correlation studies. Bitton *et al.* [30] combined proteomics data with exon array data, which allowed a much more fine-grained analysis by assigning peptides to their originating exons instead of mapping transcripts and proteins based on their IDs.

The ability of kernel methods to deal with complexly structured data was understood and demonstrated in 2002, when a crucial paper [197] integrated heterogeneous genomic data (that is, amino acid sequence information, expression data, protein-protein interaction (PPI) data, and other types of genomic information) to solve a single classification problem: the classification of transmembrane versus non-transmembrane proteins. Each data set was transformed into a kernel matrix and integration occurred on this kernel level without referring back to the data. Thanks to this integration of information, a higher accuracy was achieved than possible when based on any of the data sets separately. Because it was not known in advance whether all considered data sets were useful for protein classification and due to the amount of false positive interactions in PPI networks, a framework was required to determine the relevance of each data set, expressed as weights assigned to the kernel matrices. A convex optimization problem for simultaneously optimizing the hyperplane selection and the selection of kernel weights was proposed, solvable with semi-definite programming techniques [196].

The framework of Lanckriet has resulted in many other integration methods for the prediction of gene function and protein structure. Kato and colleagues [179] implemented a kernel-based network prediction method with automatic data selection from gene expression, phylogenetic profiles and amino acid sequences. For fold recognition, a kernel machine was proposed for the integration of heterogeneous feature groups such as protein characteristics and sequence-alignment features [90]. This method simultaneously operates in multiple feature spaces via a kernel combination methodology. Finally, Barutcuoglu and colleagues [17] improved gene function classification by combining predictions of multiple SVM classifiers in a Bayesian framework.

3.3 Data

In Europe in 2006, an estimated 3,191,600 cancer cases were diagnosed while there were 1,703,000 deaths from cancer [117]. Colon and rectal cancer are the second most common form of cancer (412,900 patients, 12.9%) and the second leading

cause of cancer-related deaths in both men and women (over 200,000 deaths, 12.2%). These numbers justify the need in Europe towards organized colorectal cancer screening programmes to stop the increasing trend in the number of deaths from this cancer. On the other hand, prostate cancer in Europe in 2006 was the most frequent cancer diagnosed in men due to the use of prostate-specific antigen tests (345,900 patients, 20.3%). Despite the widespread use of this test, prostate cancer predominantly affects elderly men, causing this disease to be the third most common form of cancer death with an increase in the number of deaths to 87,400 (9.2%).

3.3.1 Data Set I on Rectal Cancer

Compared to colon cancer, rectal cancer is a more complex disease. It is not only associated with a significant risk of systemic spread, but also with a significant risk of local relapse. The standard of care is preoperative radiotherapy with or without chemotherapy [33].

Patients and Treatment

Forty patients with rectal cancer (T_3-T_4 and/or N+) from seven Belgian centers were enrolled in a phase I/II study investigating the combination of cetuximab, capecitabine and external beam radiotherapy in preoperative treatment [218]. These patients received preoperative radiotherapy (1.8 Gy, 5 days/week for 5 weeks) in combination with cetuximab (initial dose, 400 mg/m^2 intravenous, 1 week before the beginning of radiation; followed by 250 mg/m^2/week for 5 weeks) and capecitabine for the duration of radiotherapy (first dose level, 650 mg/m^2 orally twice-daily; second dose level, 825 mg/m^2 twice-daily; including weekends). Details of the eligibility criteria, pretreatment evaluation, radio- and chemotherapy, cetuximab administration, surgery, follow-up and histopathological assessment of response to chemoradiation have been published [218].

Data Preprocessing

Tissue and plasma samples were gathered at three time points: before treatment (T_0), after the first loading dose of cetuximab but before the start of radiotherapy with capecitabine (T_1), and at the moment of surgery (T_2). All experimental procedures were done following standard laboratory procedures or the manufacturers' instructions. Because of the exclusion of some patients due to a missing outcome value, death before surgery or not having had surgery, the data set ultimately contained 36 patients.

The frozen tissue samples were hybridized to Affymetrix human U133 2.0 plus gene chip arrays. The resulting data were first preprocessed for each time point separately using robust multichip analysis (RMA) [169]. Secondly, the number of features was reduced from 54,613 probe sets to 27,650 genes by taking the median of all probe sets that matched the same gene. Probe sets that matched multiple genes were excluded because of the danger of cross-hybridization. Taking into account the low signal-to-noise ratio of microarray data, we finally filtered out genes with low variation across all samples. Only retaining the genes with a variance in the top 25% reduced the number of features to 6,913 genes.

Ninety-six proteins known to be involved in cancer were measured in the plasma samples using a Luminex 100 instrument. Proteins that had absolute values above the detection limit in less than 20% of the samples were excluded for each time point separately. This resulted in the exclusion of six proteins at T_0, four at T_1 and six at T_2. The proteomics expression values of transforming growth factor alpha, which had too many values below the detection limit, were replaced by the results of ELISA tests performed at the Department of Experimental Oncology in Leuven, Belgium. For the remaining proteins the missing values were replaced by half of the minimum detected for each protein over all samples, and values exceeding the upper limit were replaced by the upper limit value. Because most of the proteins had a positively skewed distribution, a log transformation (base 2) was performed.

In this chapter, only the data sets at T_0 and T_1 were used because our goal was to predict the different outcomes described in the next section before therapy or early in therapy.

Response Classification

A semiquantitative classification system was described by Wheeler *et al.* [362] to determine histopathological tumor regression (that is, the responsiveness of the tumor to preoperative therapy). On the other hand, two important prognostic factors in rectal cancer are pathologic lymph node involvement and CRM [217]. Because the completeness of tumor resection relies on the assessment of resection margins by the pathologist, knowledge of CRM before therapy provides important information about local recurrence and development of distant metastasis and survival [1].

For these three outcomes, registered for 36 patients at the moment of surgery, *responders* were distinguished from *nonresponders*. The grading of regression established by Wheeler and colleagues [362] (from now on referred to as WHEELER) is a modified pathological staging system for irradiated rectal cancer. It includes a measurement of tumor response after preoperative therapy: grade 1, good responsiveness (tumor is sterilized or only microscopic foci of adenocarcinoma remain); grade 2, moderate responsiveness (marked fibrosis with still a macroscopic

tumor); and grade 3, poor responsiveness (little or no fibrosis with abundant macroscopic tumor). Tumors were classified as responder when assigned to grade 1 (26 patients) and nonresponder when assigned to grade 2 or 3 (10 patients). Response can also be evaluated with the pathologic lymph node stage at surgery (pN-STAGE). The responder class contained 22 patients with no lymph nodes found at surgery while the nonresponder class contained 14 patients with at least 1 regional lymph node. CRM was measured according to the guidelines of Quirke *et al.* [260]. CRM was considered positive when the distance between the tumor and the mesorectal fascia was $\leq$2mm. Tumors with a negative CRM were labeled as responder (27 patients), while tumors with a positive CRM belonged to the nonresponder class (9 patients). Thirteen patients belonged to the responder class for all three outcomes, while there was an overlap of two patients between the nonresponder classes.

Table 3.1: Overview of the two case studies on rectal and prostate cancer

	Data set I (rectal cancer)	Data set II (prostate cancer)
Number of samples	36	55
Data sources	Transcriptomics Proteomics	Transcriptomics Genomics
Number of features (after preprocessing)	T_0: 6,913 genes; 90 proteins T_1: 6,913 genes; 92 proteins	6,974 genes 7,305 CNVs
Outcomes	WHEELER (26 - 10) pN-STAGE (22 - 14) CRM (27 - 9)	GRADE (36 - 19) STAGE (25 - 25) METASTASIS (38 - 12) RECURRENCE (22 - 7)

3.3.2 Data Set II on Prostate Cancer

We also applied our method to a publicly available data set on prostate cancer [198] [199].

Patients and Treatment

Lapointe and colleagues [199] first profiled gene expression in 71 prostate tumor cases of which 62 were primary and 9 had lymph node metastases. A cDNA microarray was used, containing 39,711 human cDNAs representing 26,260 mapped

genes. Additionally, CNVs were profiled on cDNA microarrays for CGH in 64 prostate tumor cases, among which 55 were primary tumors [198]. The arrays were obtained from the Stanford Functional Genomics Facility and included 39,632 human cDNAs corresponding to 22,279 genes. All tumors were removed by radical prostatectomy (that is, the surgical removal of the prostate gland). Among the primary tumors, the available gene expression and genomics data were in common for 55.

Data Preprocessing

Median fluorescence ratios were calculated for genes represented by multiple arrayed cDNAs. Missing gene expression values were imputed unsupervised using the s-nearest neighbors method of Troyanskaya so that a missing value for a spot S in a sample was estimated as the weighted average of the 15 spots most similar to spot S in the remaining samples. The same unsupervised prefiltering as applied to the rectal cancer data set was used for both the microarray and genomics data set. Features with a variance in the top 50% were retained, reducing the data sets to 6,974 genes and 7,305 CNVs, respectively.

Response Classification

The two pathological variables stage and grade, metastasis of the tumor and the outcome after prostatectomy defined as recurrence were considered. For grade (from now on referred to as GRADE), the Gleason Grading system was used, which is based on the most and second most common architectural patterns in the glands of the tumor [141]. Two groups could be distinguished based on the architecture of the pattern: 36 tumors were well differentiated (that is, low-grade), 19 were poorly differentiated (that is, high-grade). According to the extent of the primary tumor (STAGE), 25 samples were of stage T_2 (that is, the cancer is confined within one lobe of the prostate gland), while 25 samples were of advanced stage T_3 (that is, the tumor has extended through the fibrous tissue surrounding the prostate gland but no other organs are affected). The stage of the remaining five patients was not known. The cancer had metastasized to distant lymph nodes in 12 tumors, while the cancer had not spread beyond the regional lymph nodes in 38 of the tumors (METASTASIS). Finally, tumor recurrence was defined as a rise in prostate-specific antigen of at least 0.07 ng/ml or as occurrence of clinical metastasis (RECURRENCE). Seven tumors recurred while 22 tumors did not. The recurrence status for the remaining 26 patients was not available.

3.4 Methodology

In this section, an overview is given of the integration approach and model building strategy. Some remarks are made at the end of the section on the choice of the weights assigned to the different data sets.

3.4.1 Data Fusion

Three ways exist for kernel methods to learn simultaneously from multiple data sources. They differ in the stage of the model building process at which integration occurs and are referred to as early, intermediate and late integration [244]. Figure 3.1 gives a global overview of these three methods in case of two available data sets. *Early integration* refers to integration at database level. This means that the data sets are concatenated before a model is built. With *intermediate integration*, data sets are treated as separate entities and combination occurs at kernel level. Local features are first created by applying a kernel function on each data set separately, to extract relationships between variables within a data set while ignoring relationships between different types of variables. The corresponding kernel matrices are added or combined via a weighted sum (that is, $K = \sum_{m=1}^{M} \mu_m K_m$), corresponding to linearly combining the local features to create global features. The latter features are incorporated in the final model that is trained on the combined kernel matrix K. In the *late integration approach*, models are built on each of the data sets separately before combining their predictions with a decision function, to obtain a single prediction per observation.

In this dissertation, intermediate integration is chosen due to its advantage over the other two approaches [244]. With intermediate integration, the nature of each data set is better taken into account compared to early integration by adapting the kernel functions to each data set separately. In this way, relations between features of the same data set can be modeled more accurately, while constructing the decision boundary in an explicit heterogeneous feature space. On the other hand when compared to late integration, only one predicted outcome per patient is obtained by adding the kernel matrices before training the wLS-SVM. Intermediate integration relies on a single hypothesis and makes the extra decision function which is needed for late integration unnecessary. Intermediate integration thus makes a trade-off between allowing too many dependencies between data sets (in early integration) and making too many independence assumptions (in late integration). We also validated the superiority of intermediate integration numerically by comparing the three approaches on the data sets described in Section 3.3. For prostate cancer, four outcomes were predicted with the microarray and genomics data set, while for rectal cancer microarray and proteomics data were combined at T_0 and T_1 separately for the prediction of the three available outcomes,

resulting in 10 distinct cases in total. The normalized linear kernel function was used. For intermediate and late integration, weights μ_m were set equal while the regularization parameter γ was optimized with the grid search approach for each individual classifier. The AUC values obtained with the different integration strategies for all combinations of number of genes and number of proteins or CNVs, ranging from 5 to 35 in steps of 5, were compared using the two-sided paired-sampled t-test. Intermediate integration outperformed late integration for 8 out of 10 cases with p-values ranging from 5.52e-17 to 0.358 (median p-value = 2.37e-5), while for the other 2 cases the performance of late integration was not significantly better (p-values = 0.4383 and 0.6933). Intermediate integration was better than early integration in 6 cases (p-values 5.24e-30 to 7.64e-8, median = 2.59e-11). For three of the remaining cases (that is, CRM at T_1, GRADE and RECURRENCE), early integration performed significantly better at a significance level of 0.05. This may be explained by the discrepancy in predictive information contained in the individual data sets, which causes the majority of features to be selected from the most informative data set. For intermediate and late integration with equal weights assigned to both data sets, a less informative data set can deteriorate performance.

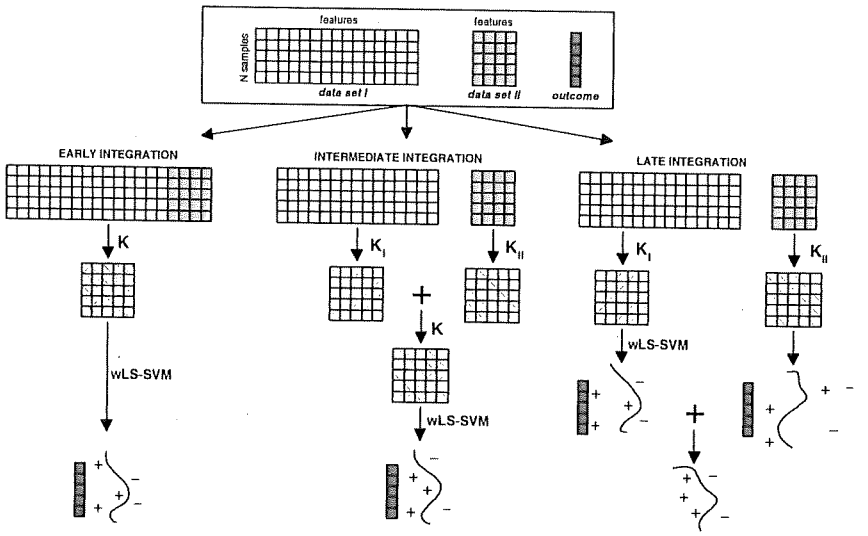


Figure 3.1: *Three methods to learn from multiple data sets:* The data sets are represented as matrices with rows corresponding to patients and columns to features. In early integration, a wLS-SVM model is trained on the kernel matrix, computed from the concatenated data set. In intermediate integration, a kernel matrix is computed for both data sets and a wLS-SVM is trained on the sum of the kernel matrices. In late integration, a wLS-SVM model is trained for each data set separately. A decision function results in a single outcome per patient.

3.4.2 Model Building

When integrating kernel matrices according to the intermediate integration approach, normalization is required to guarantee the same order of magnitude. The normalized linear kernel function in (2.3) with $k(x^i, x) = x^{i^T} x$ was used. Positive semidefiniteness of the linear combination of kernel matrices $K = \sum_{m=1}^{M} \mu_m K_m$ with M the number of data sets is guaranteed by constraining the weights μ_m to be non-negative. Due to the high-quality data gathered from the patients and the rather small sample size in both case studies, weights were chosen equally.

The feature selection method DEDS was applied to the microarray data sets and the genomics data set, whilst the Wilcoxon rank sum test was used for the proteomics data (see Section 2.3). A grid search in combination with LOO was applied for parameter selection and estimation of the generalization performance. For both case studies, a three-dimensional grid sufficed for the optimization of the regularization parameter γ, the number of genes selected from the microarray data sets, and the number of proteins or CNVs obtained from the proteomics or genomics data set, respectively. For the data on rectal cancer, the number of genes/proteins selected at T_0 and T_1 was set equal when data from both time points were considered. For each considered outcome, the AUC of the optimal performing model was compared with the AUC of the other models using the method of Hanley and McNeil [155]. A p-value of 0.05 was considered statistically significant. The final features were chosen as those that occurred most often in the top rankings determined in each LOO iteration. Figure 3.2c gives an overview of the strategy, limited to two data sets for clarity.

Two other strategies were considered for comparison with intermediate integration. This allowed verifying whether a complex integration of data over multiple layers of biological regulation is crucial. First, wLS-SVM models were built on each data set separately, mimicking the results that would have been obtained when only static data from one platform were available (Figure 3.2a). Due to one set of features and the use of a linear kernel function, a two-dimensional grid sufficed for the optimization of γ and the number of features. Secondly, when measurements are repeated at multiple time points, knowledge over time can be exploited. For rectal cancer, data were available before and early in therapy. Combination was performed for each data type separately by manually calculating the change in gene expression or protein abundance ($T_0 - T_1$). By using these changes in time as features, a two-dimensional grid also sufficed, as shown in Figure 3.2b. For all models the parameters were trained according to the same approach, which makes the corresponding LOO results comparable for each outcome separately. In developmental phase, these three strategies required 5.11 years of effective, parallelized computing time corresponding to 149,000 jobs, in order to obtain kernel-based models for therapy response prediction. Especially the grid-based

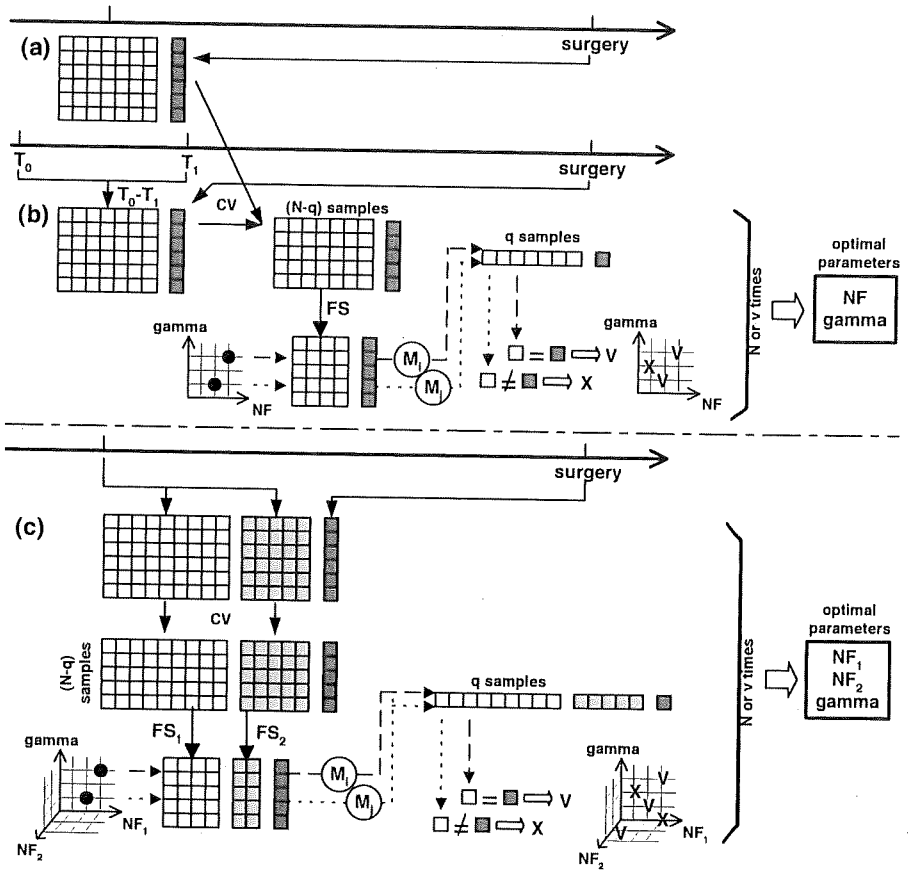


Figure 3.2: *Overview of the model building strategies with LOO ($q = 1$) or v-fold CV ($q = N/v$):* **(a)** wLS-SVM models are built on each data set separately. A two-dimensional grid is used for the optimization of the regularization parameter γ and the number of features; **(b)** Data sets of different time points are manually combined. By using the changes in expression or abundance as features, the same strategy as in (a) can be applied; **(c)** The intermediate integration method is used for a genome-wide integration of all available data sets. An r-dimensional grid is required for optimizing γ and the number of features selected from the $(r-1)$ integrated data sets. FS, feature selection; NF, number of features; T, time point; M_i, model for parameter combination i.

parameter optimization was computationally intensive.

Insights into biological functions involved in a specific cancer type can be enlarged by assessing the Gene Ontology (GO) categories [11] that are significantly over-represented in gene lists resulting from microarray, proteomics and other large-scale analyses. The web-based platform GOTM (Gene Ontology Tree Machine) [379] is a GO enrichment analysis tool. The hypergeometric test was used to assess the enrichment of each GO category with respect to the whole genome, while the resulting p-values were corrected for multiple testing using the false discovery rate correction [21].

3.4.3 Weight Optimization

The weights μ_m assigned to the kernel matrices K_m can be specified beforehand, or, since one data set possibly contains more information than other sets, the weights can be learned from data. When the data sets or kernel matrices are normalized, the optimized weights reflect the relative importance of the different data sets. Optimization of the weights is important when dealing with many data sets of which only several are relevant. A redundant data set or a data set with much noise will receive a weight close to zero and its influence on the outcome is kept small. However, when the number of data sets is limited and most of them are reliable and relevant to the problem at hand, a trade-off needs to be made between performance and computational burden (for example, extra required CV loops). Previous studies have shown that only in case of redundant or noisy data sets, optimization of weights leads to a better performance [197]. This is supported by the equal weights theorem presented by Wainer [355]. He formally proved that when all optimized weights are uniformly distributed on the interval [0.25; 0.75], use of equal weights barely changes performance. In our integration approach, all data sets (or kernel matrices) are standardized, meaning that Wainer's theorem holds when no data set is completely redundant. This is often true for patient-related data, that is, all data sets are relevant with no data set containing sufficient information for perfect classification. We therefore opted for equal weights. Moreover, optimizing weights reduces the degree of freedom available for selecting other parameters [355]. Due to the complexity of high dimensional data, part of the noise in the training data will be modeled, resulting in overfitting and an overestimate of the generalization performance. The estimates of weights are also undesirably influenced by outliers, which are not representative for validation data and lead to a reduction in accuracy. Both latter problems are avoided when using equal weights.

To support the choice of equal weights, we considered the combination of clinical and microarray data for the prediction of distant metastasis in breast cancer patients and investigated the effect of weight optimization on performance [79].

The grid was extended with one dimension for the optimization of μ assigned to clinical data. The weight assigned to microarray data was defined as $1 - \mu$. The weights were optimized with a linesearch by increasing the weight of one kernel from 0 to 1 in steps of 0.02 and decreasing the weight of the other kernel equally. When maximizing sensitivity and keeping specificity as high as possible, a weight of 0.28 was assigned to the clinical data set and 0.72 to the microarray data set, resulting in a model with an AUC of 0.66 ($\pm$0.049). No significant difference was observed when equal weights of 0.5 were *a priori* assigned to both data sets (AUC = 0.662$\pm$0.049).

3.5 Toolbox HI-DID-IT

At our institution, two LS-SVM related tools have already been developed. LS-SVMlab [246] is a Matlab/C toolbox containing a variety of techniques and algorithms for LS-SVMs, applicable to nonlinear regression, function estimation and classification. Fixed size LS-SVMs for large data sets, extensions towards robustness, sparseness and weighted versions, and different techniques for tuning hyperparameters are included. While this toolbox is suited for researchers and practitioners in LS-SVMs in general, the web service M@CBETH, a MicroArray Classification BEnchmarking Tool on a Host server [255], was made available for two-class microarray prediction tasks. It aims at finding the best prediction among different classification methods by using randomizations of the benchmarking data set. Both tools, however, are restricted to the analysis of a single data set. As the fusion of various types of genome-wide data is an inevitable requirement to improve prognosis, diagnosis and therapy response prediction, we have expanded our data integration framework from a methodological perspective to the training and testing of models with clinical applicability [84]. HI-DID-IT (HIgh-DImensional Data Integration Toolbox) is a software implementation in Matlab through which multiple high-dimensional data sets can be integrated for clinical decision support. This toolbox is freely available under the GNU GPL licence at http://homes.esat.kuleuven.be/~bioiuser/HIDIDIT/index.php.

The toolbox HI-DID-IT can be used for comparison with other methodological approaches in benchmarking studies, for model training as well as model testing before being used in clinical trials. The LS-SVM is used for classification and regression. Different integration schemes, the optimization of the weights assigned to the different data sets and feature selection strategies are implemented. For five extra issues, a certain degree of freedom is offered to the user.

Integration schemes: The methods early, intermediate and late integration are implemented to simultaneously learn from multiple data sets.

Weight optimization: In specific applications, the choice of the weights μ_m may

be important. The users are therefore given the possibility to decide in function of their specific problem at hand whether weight optimization may be beneficial.

Feature selection: The toolbox allows building a model on a predefined number of features, selected with the Wilcoxon rank sum test for classification or the Spearman correlation coefficient for regression. On the other hand, the number of features can be optimized according to a specified criterion. Due to the clinical advantage of models based on a limited set of features, the number of features is subsequently calculated at which performance has dropped with a predefined percentage with respect to performance at the optimal number of features.

Kernel functions: In the toolbox, the kernel functions that are most frequently employed in classification problems are available: the linear, polynomial and RBF kernel (see Section 2.2). The specific choice of the kernel function depends on the distribution of the data entities. As required for data integration, the normalized version (2.3) of these kernel functions is also provided.

Tuning parameter optimization: The parameters that have to be optimized are the number of features, the weights assigned to the different data sets and the parameters of the kernel method, being γ, the intercept constant τ and degree d for the polynomial kernel, and width σ for the RBF kernel. Within each performance loop, an internal LOO or 5-fold CV strategy is applied to the training data for the optimization of all parameters. For the selection of γ, 40 default values are considered. The parameters τ and d vary on a linear scale, with 10 default values for τ, and d varying from 1 to 5. For the RBF kernel, 10 default values for σ are defined on an exponentially growing scale. A locally refined grid for the parameters γ, τ and σ is defined around their optimized values to further fine-tune them. For this second grid, the same number of parameter values are considered as for the first grid.

Performance assessment: When only training data are available, it is possible either to use all data at once for the optimization of the parameters, or to use a certain type of cross-validation to reduce the risk of overfitting (that is, LOO for small data sets; v-fold CV or randomized splits of data into a training and test part for larger data sets). The default setting is 5-fold CV for a sample size < 50 and 10-fold CV otherwise, both with stratification of the samples to outcome. When randomization is selected by the user without further specifications, 10 stratified randomizations with 70% of the data used for training are performed. A set of independent observations on which the trained model is validated can also be provided.

Optimization criteria: The parameters are optimized according to different criteria, with the choice of the criterion depending on the problem at hand. For classification problems, the AUC is most often maximized, although it is possible to minimize the BER or classification error, or to maximize sensitivity or the Youden's index. For regression problems, the parameters are chosen corresponding

to the model with the lowest MSE.

Weighted LS-SVM / stratification: Specific to classification problems, a weighted version of the LS-SVM is implemented to account for the unbalance in cancer data sets (see Section 2.2.4). Moreover, the option is provided to split the observations with stratification to outcome, to ensure that the relative proportion of observations sampled in both training and test set is similar to the original proportion in the full data set.

3.6 Results

After the methodological aspects of data integration, this section provides the results for both considered case studies. Our method has also been compared with an ensemble approach, and the results for both a generalizability and correlation analysis are presented.

3.6.1 Study I on Rectal Cancer

Using the methodologies shown in Figure 3.2, models were built on microarray and proteomics data of 36 rectal cancer patients at two time points during therapy, for the prediction of three outcomes registered at the moment of surgery: a tumor regression grading system (WHEELER) and two prognostic factors, pathologic N stage at surgery (pN-STAGE) and the circumferential resection margin (CRM). The models obtained with our integration approach comprise MT_{01} (model based on microarray data at both time points), PT_{01} (model based on proteomics data at both time points), MPT_0 (model based on microarray and proteomics data at T_0), MPT_1 (model based on microarray and proteomics data at T_1), and all possible combinations of data sets for which both technologies and a sample from both time points are required (using the same name convention). To investigate the advantage of our integration approach, we additionally investigated the models MT_0, MT_1, PT_0 and PT_1 based on one data set, as well as the models based on the change in gene expression or protein abundance between T_0 and T_1 (that is, $MT_0 - T_1$ and $PT_0 - T_1$). The number of genes and proteins were chosen to optimize the LOO performance of the wLS-SVM models. The features selected most often in the 36 LOO iterations are listed in Table 3.3 and discussed later.

Table 3.2 shows the wLS-SVM models for the considered combinations of data sets to predict WHEELER, pN-STAGE and CRM, with the number of genes and proteins that resulted in the highest AUC and Youden's index and lowest BER. Genes and proteins were selected with DEDS and the Wilcoxon rank sum test, respectively. The corresponding ROC curves are shown in Appendix A. Combining microarray and proteomics data at T_1 resulted for all three outcomes into an

increased performance with respect to models based on one data set or on the change in gene expression or protein abundance. The p-values for the comparison of the ROC curve of MPT_1 with those of all other models are provided in Table 3.2 [155]. For the prediction of WHEELER, the expression of 25 genes and 12 proteins at T_1 was better, although not significantly, with a LOO AUC of 0.9269. For pN-STAGE, combining both data sets at T_1 using the expression of 21 genes and 14 proteins resulted in a significantly better AUC of 0.9870 compared to models based on one data set or on the change in expression or abundance over time. Finally, the inclusion of 7 genes and 33 proteins at T_1 led to an AUC of 0.9630 for the prediction of CRM, with 4 models based on only one data type performing significantly worse compared to MPT_1. For all three outcomes, one model requiring both technologies and a sample to be taken at two time points performed even better than MPT_1, although not significantly (that is, MT_0PT_1 for WHEELER, MT_0PT_{01} for pN-STAGE and $MT_{01}PT_1$ for CRM). Due to the limited number of samples in this study, however, no conclusion could be drawn on whether the effort of two technologies and two samples is needed. Larger studies are required for a confident trade-off between cost-efficiency and accuracy. Finally, none of the selected genes were transcripts of the selected proteins.

Table 3.2: wLS-SVM models for the prediction of WHEELER, pN-STAGE and CRM in rectal cancer

	Model	NG$^{\mu}$	NP$^{\varsigma}$	AUC (SE)$^{\varphi}$	p-value$^{\varrho}$
WHEELER					
	MT_0	4		0.7538 (0.1085)	0.0987
	MT_1	29		0.9038 (0.0502)	0.6861
	PT_0		35	0.7423 (0.0867)	0.0540
	PT_1		11	0.9038 (0.0575)	0.7273
	$MT_0 - T_1$	32		0.6846 (0.1215)	0.0598
	$PT_0 - T_1$		5	0.8654 (0.0621)	0.4135
	MT_{01}	3^{π}		0.7808 (0.0985)	0.1320
	PT_{01}		21^{π}	0.7692 (0.0831)	0.0831
	MPT_0	3	35	0.8461 (0.0718)	0.2760
	MPT_1	**25**	**12**	**0.9269 (0.0425)**	
	MT_0PT_1	2	4	0.9385 (0.0444)	0.8101^{ϕ}
	MT_1PT_0	26	1	0.9154 (0.0471)	0.8304
	$MT_{01}PT_0$	2^{π}	32	0.8885 (0.0544)	0.5398
	$MT_{01}PT_1$	2^{π}	3	0.9154 (0.0490)	0.8365
	MT_0PT_{01}	4	3^{π}	0.8808 (0.0597)	0.4765
	MT_1PT_{01}	28	1^{π}	0.8885 (0.0544)	0.4931

Continued on next page

Table 3.2 – continued from previous page

	Model	NG$^{\mu}$	NP$^{\varsigma}$	AUC (SE)$^{\varphi}$	p-value$^{\varrho}$
	MPT_{01}	2^{π}	31^{π}	0.8846 (0.0558)	0.4858
pN-STAGE					
	MT_0	25		0.6493 (0.0914)	2.315e-4
	MT_1	22		0.8506 (0.0665)	0.0362
	PT_0		2	0.6753 (0.0906)	6.659e-4
	PT_1		12	0.8409 (0.0652)	0.0238
	$MT_0 - T_1$	4		0.6071 (0.0986)	1.359e-4
	$PT_0 - T_1$		9	0.7662 (0.0900)	0.0153
	MT_{01}	24^{π}		0.9286 (0.0450)	0.1998
	PT_{01}		34^{π}	0.8182 (0.0695)	0.0145
	MPT_0	27	27	0.9188 (0.0469)	0.1591
	MPT_1	**21**	**14**	**0.9870 (0.0135)**	
	MT_0PT_1	31	22	0.9643 (0.0261)	0.4027
	MT_1PT_0	22	2	0.9091 (0.0477)	0.1042
	$MT_{01}PT_0$	24^{π}	2	0.9286 (0.0441)	0.1729
	$MT_{01}PT_1$	31^{π}	14	0.9675 (0.0249)	0.4417
	MT_0PT_{01}	26	20^{π}	1 (0)	0.3347^{ϕ}
	MT_1PT_{01}	10	17^{π}	0.9805 (0.0178)	0.7595
	MPT_{01}	23^{π}	16^{π}	0.9610 (0.0280)	0.3421
CRM					
	MT_0	33		0.6790 (0.1016)	0.0072
	MT_1	9		0.9259 (0.0472)	0.4955
	PT_0		34	0.8518 (0.0624)	0.0935
	PT_1		34	0.7654 (0.0831)	0.0281
	$MT_0 - T_1$	6		0.9136 (0.0480)	0.4030
	$PT_0 - T_1$		2	0.8272 (0.0709)	0.0849
	MT_{01}	16^{π}		0.8066 (0.0846)	0.0468
	PT_{01}		3^{π}	0.7531 (0.0865)	0.0227
	MPT_0	7	27	0.8477 (0.0688)	0.1340
	MPT_1	**7**	**33**	**0.9630 (0.0344)**	
	MT_0PT_1	6	27	0.8230 (0.0841)	0.1234
	MT_1PT_0	16	14	0.9630 (0.0376)	1
	$MT_{01}PT_0$	14^{π}	3	0.8807 (0.0589)	0.0975
	$MT_{01}PT_1$	9^{π}	29	0.9876 (0.0146)	0.4924^{ϕ}
	MT_0PT_{01}	2	3^{π}	0.7942 (0.0843)	0.0638
	MT_1PT_{01}	26	10^{π}	0.9424 (0.0381)	0.6794
	MPT_{01}	2^{π}	3^{π}	0.8230 (0.0771)	0.0973

$^{\mu}$Number of genes selected in each LOO iteration. $^{\varsigma}$Number of proteins selected in each LOO iteration. $^{\pi}$Number of features used at both time points. $^{\varphi}$Area under the ROC curve (standard error) obtained with leave-one-out. $^{\varrho}$Comparison of the AUC of each model with MPT_1 in bold [155]. $^{\phi}$This model is better than the model in bold we compared it with.

The genes and/or proteins that were most frequently selected in the LOO iterations of MPT_1 and contribute to rectal or colorectal cancer are shown in Table 3.3. A protein important for CRM, for example, is the epidermal growth factor receptor (EGFR), involved in signaling pathways affecting cellular growth, differentiation and proliferation. This protein represents one of the most promising targets allowing progress in colorectal cancer treatment. It has been suggested that EGFR polymorphisms as well as polymorphisms of other genes active in the EGFR pathway may be potential indicators of radiosensitivity in patients with rectal cancer treated with chemoradiation [381]. In colorectal cancer, pro-inflammatory cytokines such as interleukin-1 beta and interleukin-6 may be accountable for the overexpression of *COX-2*, important not only in early stage but also for progression [220]. Transforming growth factor alpha, down-regulated in our patients with a good responsiveness to preoperative therapy, is implicated in metastatic spread of colon cancer cells [282]. The expression of interleukin-8 is associated with induction and progression of colorectal carcinoma and the development of colorectal liver metastases [275]. In our data set, it is down-regulated in the group of patients with no lymph nodes found at surgery. Finally, elevated carcinoembryonic antigen and cancer antigen 19-9 are related to poor outcome in colorectal cancer [210]. Their levels are low in patients with no lymph nodes, while carcinoembryonic antigen is also less expressed in patients with a negative CRM, that is, belonging to the class of responders. We refer to additional data file 3 in [83] for a complete list of the genes and proteins chosen by the models MPT_1.

The predictions seem to depend on mainly different subsets of features. The gene encoding PAI-2 is important for both WHEELER and CRM, while the proteins important for two of the three outcomes are interleukin-4, ferritin, apolipoprotein H, epidermal growth factor, matrix metalloproteinase-2 and lymphotactin. Notably, these genes and proteins were also selected by the other models based on microarray and/or proteomics data at T_1, and some of them were included in the models based on data at T_0. Besides studying the contribution of each gene in cancer individually, we determined the GO categories that were over-represented in the sets of genes and proteins included in the MPT_1 models. The biological processes that all three outcomes have in common are apoptosis and the regulation thereof, response to external and chemical stimuli, and regulation of the developmental process, phosphorylation, the protein modification process and cell proliferation. An important molecular function is receptor binding.

Table 3.3: Features for (colo)rectal cancer selected by MPT_1 and known to be involved in this type of cancer

	Gene/protein	Hits$^{\varphi}$	Region	Function	↑/↓$^{\varsigma}$	Reference
W						
	COX-2	36	1q25.2-q25.3	Progression	↑	[220]
	IL-1B	36	2q14	Inflammatory response	↑	[220]
	Ferritin	36	11q13; 19q13.3-q13.4	Iron storage	↓	[262]
	EGF	36	4q25	Cell growth/proliferation/differentiation	↑	[62]
	MMP-2	36	16q13-q21	Invasion/metastasis	↑	[181]
	TGFα	36	2p13	Angiogenesis/cell proliferation	↓	[282]
	SELE	25	1q22-q25	Progression/metastasis	↑	[328]
	GM-CSF	24	5q31.1	Maintenance of granulocytes/macrophages	↑	[110]
	MMP-1	15	11q22.3	Tumor invasion/metastasis/poor prognosis	↑	[384]
N						
	REG4	36	1p13.1-p12	Early carcinogenesis	↓	[382]
	MUC2	36	11p15.5	Deregulated by TNFα	↓	[4]
	CA1	36	8q13-q22.1	Carbonate dehydratase activity	↓	[192]
	CA2	36	8q22	Carbonate dehydratase activity	↓	[192]
	CLDN8	36	21q22.11	Tumorigenesis	↓	[148]
	CEA	36	19q13.1-q13.2	Cell adhesion; tumor marker for recurrence	↓	[210]
	IL-1ra	36	2q14.2	Carcinogenesis	↑	[352]
	CA19-9	36		Tumor marker for recurrence	↓	[210]
	Ferritin	36	11q13; 19q13.3-q13.4	Iron storage	↓	[262]
	IL-1beta	36	2q14	Inflammatory response	↓	[220]
	Beta2-microglobulin	36	15q21-q22.2	Metastasis	↑	[185]
	RARRES1	31	3q25.32-q25.33	Cell proliferation	↓	[375]
	IL-8	28	4q13-q21	Progression/metastasis	↓	[275]

Continued on next page

Table 3.3 – continued from previous page

	Gene/protein	Hits[φ]	Region	Function	↑/↓[ς]	Reference
	TNFRII	24	1p36.3-p36.2	Apoptosis	↑	[233]
C						
	ICAM-1	36	19p13.3-p13.2	Metastasis	↓	[219]
	CEA	36	19q13.1-q13.2	Cell adhesion; tumor marker for recurrence	↓	[210]
	MMP-2	36	16q13-q21	Invasion/metastasis	↑	[181]
	Adiponectin	36	3q27	Metabolic/hormonal processes	↓	[118]
	Thrombospondin-1	36	15q15	Angiogenesis/tumor growth	↑	[231]
	EGF-R	36	7p12	Cell growth/proliferation/differentiation	↑	[381]
	Tissue factor	35	1p22-p21	Angiogenesis/metastasis	↑	[356]
	CYP1B1	35	2p21	Drug metabolism	↓	[25]
	EGF	32	4q25	Cell growth/proliferation/differentiation	↑	[62]

W, WHEELER; N, pN-STAGE; C, CRM. [φ]Number of occurrences of the gene/protein in the 36 LOO iterations.
[ς]Up-regulation (↑) or down-regulation (↓) in the good responders with respect to moderate or poor responders; no lymph nodes with respect to at least one regional lymph node; negative CRM with respect to positive CRM.

3.6.2 Study II on Prostate Cancer

The same methodology was applied to microarray and genomics data of 55 patients with prostate cancer. Table 3.4 shows the results for the prediction of the grade and stage of the tumor (GRADE and STAGE), the tumors that metastasized to distant lymph nodes (METASTASIS) and the tumors that recurred (RECURRENCE). The model based on both microarray and genomics data (that is, MG) was compared with model M based on microarray data and model G based on genomics data. Also in this case, after having optimized the essential number of features to be included using LOO, the final genes and CNVs were selected based on their position and number of occurrences in the 55 LOO rankings.

We obtained similar results for prostate cancer as for rectal cancer. Combining gene expression with measurements at the DNA level (MG) led, for all four outcomes, to an improvement in classification accuracy and was significant in some cases (Table 3.4). For the prediction of GRADE, six genes and eight CNVs selected with DEDS resulted in an AUC of 0.9006. For STAGE, 42 genes and 22 CNVs were needed for a performance of 0.8528. The model MG for the prediction of METASTASIS had an AUC of 0.9868 when fusing the expression of 18 genes with 3 CNVs. Finally, the prediction of RECURRENCE was most difficult, with an AUC of 0.7857 when combining 32 genes and 2 CNVs. Appendix A shows the ROC curves of the models listed in Table 3.4.

Table 3.4: wLS-SVM models for the prediction of GRADE, STAGE, METASTASIS and RECURRENCE in prostate cancer

	Model	NG$^{\mu}$	NC$^{\varsigma}$	AUC (SE)$^{\varphi}$	p-value$^{\varrho}$
GRADE					
	M	24		0.8304 (0.0623)	0.2727
	G		8	0.7822 (0.0632)	0.0503
	MG	**6**	**8**	**0.9006 (0.0413)**	
STAGE					
	M	18		0.6576 (0.0778)	0.0191
	G		32	0.7936 (0.0631)	0.3466
	MG	**42**	**22**	**0.8528 (0.0550)**	
METASTASIS					
	M	18		0.9759 (0.0178)	0.4392
	G		12	0.8114 (0.0755)	0.0166
	MG	**18**	**3**	**0.9868 (0.0121)**	
RECURRENCE					

Continued on next page

Table 3.4 – continued from previous page

Model	NG$^{\mu}$	NC$^{\varsigma}$	AUC (SE)$^{\varphi}$	p-value$^{\varrho}$
M	24		0.7208 (0.0936)	0.5392
G		26	0.4481 (0.1433)	0.0354
MG	**32**	**2**	**0.7857 (0.0934)**	

$^{\mu}$Number of genes selected in each LOO iteration. $^{\varsigma}$Number of CNVs selected in each LOO iteration. $^{\varphi}$Area under the ROC curve (standard error) obtained with leave-one-out. $^{\varrho}$Comparison of the AUC of each model with the best model in bold [155].

Several genes and CNVs have been selected by MG and are known to be involved in prostate cancer (Table 3.5). The gene *ALOX15B* is a suppressor of prostate tumor development [28] and it is in this data set down-regulated in tumors that recurred and are of high-grade. Both *SFRP4* and *CXCL14* are inhibitors of prostate tumor growth [160] [286]. *SFRP4* is up-regulated in tumors of high-grade and *CXCL14* in tumors of advanced stage. A small deletion involving chromosomal band 21q22.3 fuses all coding exons of *ERG* to androgen-related sequences in the promoter of the prostate-specific *TMPRSS2* gene. This chromosomal rearrangement is a highly prevalent oncogenic alteration in prostate tumor cells and leads to an aberrant expression of the *ERG* proto-oncogene, important for early prostate carcinogenesis [129]. In this data set, *ERG* is overexpressed in tumors in which the cancer metastasized to distant lymph nodes. It has been shown that this genetic biomarker is a strong prognostic factor for disease recurrence, and can be used for early detection in prostate cancer [234]. *VAV3*, an oncogene involved in the development and progression of prostate cancer, is up-regulated in tumors that metastasized [100]. It has previously been shown that strong overexpression of *TIAM1* is significantly associated with disease recurrence and a decreased disease-free survival [111]. *JAG1*, also significantly associated with recurrence [280], plays a role in cell growth, progression and metastasis. In this data set, both genes are up-regulated in the group of tumors that recurred. Finally, several germline mutations or variants in *RNASEL* have been observed among hereditary prostate cancer cases, indicating that polymorphic changes within the *RNASEL* gene may be associated with increased risk of familial prostate cancer [300]. A list of all the genes and CNVs selected by the models MG can be found in the additional data file 3 in [83]. The outcomes for prostate cancer are also characterized by mainly different sets of features. Five genes overlap between at least two outcomes (*ERG*, *AHSG*, *SEMA4G*, *F5* and *ALOX15B*), while the same holds for four CNVs of the genes *GPD1L*, *KCTD12*, *SMYD5* and *TRO*. Important GO categories for GRADE are cell differentiation and cellular development, while the genes for prediction of METASTASIS are involved in negative regulation of cell growth.

Table 3.5: Features for prostate cancer selected by *MG* and known to be involved in this type of cancer

	Gene/CNV	Hits$^{\varphi}$	Region	Function	↑/↓$^{\varsigma}$	Reference
G						
	SFRP4	55	7p14.1	Inhibitor of PT growth/invasion	↑	[160]
	VCAN	55	5q14.3	Contributor to PC pathology	↑	[73]
	ALOX15B	36	17p13.1	Suppressor of PT development	↓	[28]
S						
	MAGEA4	50	Xq28	Only expressed in PC (diagnosis and therapy)	↓	[165]
	ANPEP	50	15q25-q26	PT cell invasion	↓	[170]
	POU4F1	50	13q31.1	PC cell growth	↓	[99]
	CXCL14	48	5q31	Inhibitor of PT growth	↑	[286]
	RNASEL	48	1q25	Polymorphic changes as tumor; suppressor in hereditary PC	↑	[300]
	GDEP	41	4q21.1	Prostate-specific gene	↓	[72]
M						
	ERG	50	21q22.3	Proto-oncogene; early prostate carcinogenesis	↑	[129]
	AREG	49	4q13-q21	PC progression/growth via *TARP*	↓	[367]
	VAV3	49	1p13.3	Oncogene; PC development/progression	↑	[100]
	ADAMTS1	26	21q21.2	Negatively effected by *TGFβ1*, which increases *VCAN*-expression	↓	[73]
R						
	AZGP1	29	7q22.1	Inversely associated to tumor stage; predictor of biochemical recurrence	↓	[96]
	TIAM1	29	21q22.1-11	Predictor of decreased disease-free survival/ recurrence	↑	[111]

Continued on next page

Table 3.5 – continued from previous page

Gene/CNV	Hits[φ]	Region	Function	↑/↓[ς]	Reference
FGG	28	4q28	PC cell growth	↓	[278]
ATF3	26	1q32.3	Inversely related to invasion/angiogenesis; positively correlated to metastasis	↓	[15]
JAG1	26	20p12.1-20p11.23	Cell growth/progression/metastasis	↑	[280]
ERG	14	21q22.3	Proto-oncogene; early prostate carcinogenesis	↑	[129]
ALOX15B	14	17p13.1	Suppressor of PT development	↓	[28]

G, GRADE; S, STAGE; M, METASTASIS; R, RECURRENCE. [φ]Number of occurrences of the gene/CNV in all LOO iterations (number of LOO iterations for G=55, S=50, M=50, R=29). [ς]Up-regulation (↑) or down-regulation (↓) in high-grade with respect to low-grade; advanced stage with respect to early stage; metastasis with respect to no metastasis; recurrence with respect to no recurrence. PC, Prostate Cancer; PT, Prostate Tumor.

3.6.3 Comparison with an Ensemble Method

The late integration approach introduced in Section 3.4.1 is a standard ensemble method. It has been shown that ensemble based systems lead to a better performance compared to single-expert systems for a wide range of applications, under a variety of scenarios, and using different ensemble based algorithms and combination rules [256]. For data fusion, each data set gives rise to a separate classifier before combining the predictions of these models. We specifically assessed the benefit of our kernel-based integration approach over late integration in which the continuous predictions were added. Table 3.6 shows the AUC values of the optimal models obtained with our strategy (that is, MPT_1 and MG) and the corresponding AUC values for late integration with the inclusion of the same number of genes, proteins and CNVs. The p-values of the significance tests comparing the ROC curves are reported as well [155]. For CRM, our strategy was significantly better than the ensemble approach at a significance level of 0.05. For WHEELER and pN-STAGE, the AUC values did not differ significantly. For prostate cancer, the AUC of MG was better than the AUC of the ensemble models, although being significantly better for RECURRENCE only.

Table 3.6: Comparison of our kernel-based integration approach with late integration

Outcome	AUC (SE)$^{\mu}$: MPT_1/MG	AUC (SE)$^{\mu}$: late integration	p-value$^{\varsigma}$
WHEELER	0.9269 (0.0425)	0.9500 (0.0339)	0.6160
pN-STAGE	0.9870 (0.0135)	0.9253 (0.0432)	0.1422
CRM	0.9630 (0.0344)	0.7860 (0.0783)	**0.0384**
GRADE	0.9006 (0.0413)	0.8567 (0.0521)	0.3745
STAGE	0.8528 (0.0550)	0.8304 (0.0582)	0.6836
METASTASIS	0.9868 (0.0121)	0.9452 (0.0309)	0.1313
RECURRENCE	0.7857 (0.0934)	0.4545 (0.1352)	**0.0182**

$^{\mu}$Area under the ROC curve (standard error) obtained with leave-one-out. $^{\varsigma}$Comparison of the AUC of the optimal models obtained with our strategy (MPT_1 for rectal cancer, MG for prostate cancer) and the corresponding ensemble models based on the same number of features [155].

3.6.4 Generalizability Analysis

Whereas in an ideal case independent test set results should be obtained, the sample size in both presented cancer studies is rather small. To support our choice of using the full data set, we verified whether the above presented

LOO results provide an appropriate view on model generalization by randomly dividing each data set into a training and test set, using the training set for model construction and validating this model on the test set. We repeated this 20 times for each outcome and all considered models. The internal LOO performances and the corresponding test performances were compared using a boxplot representation across all randomizations. Figures 3.3 and 3.4 show this comparison for WHEELER and GRADE, respectively, while we refer to Appendix B for a more detailed description and the boxplots for the other outcomes. The boxplots are wider for the test results due to the small number of test samples and the unbalance with only one or two of the test samples belonging to the minority class. These boxplots illustratively show that the rank of the models' performance is comparable for the LOO results on the full data set, the internal LOO results obtained with randomization, and the test results. Furthermore, the features chosen in each of the 20 randomizations were concordant, on average being selected in 71% of the randomizations (median 70%, interquartile range 50-95%). These findings support our motivation to present the results obtained on the full data.

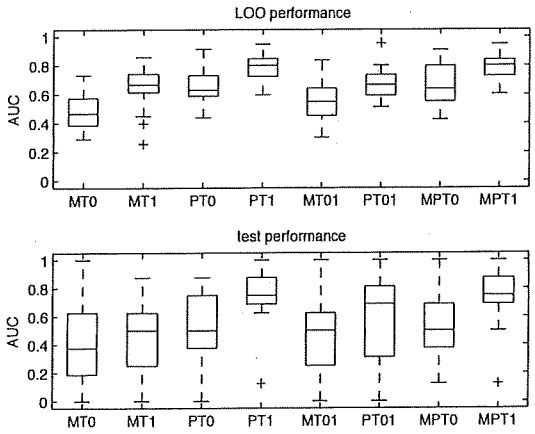


Figure 3.3: *Comparison for WHEELER of the internal LOO performances (AUC) in subplot 1 and the corresponding test performances in subplot 2 using a boxplot representation across all 20 randomizations:* Boxplots are shown for the optimal number of selected features, ranging from 1 to 35 in steps of 5: MT_0 (30 genes), MT_1 (25 genes), PT_0 (30 proteins), PT_1 (10 proteins), MT_{01} (5 genes), PT_{01} (15 proteins), MPT_0 (5 genes, 35 proteins), and MPT_1 (1 gene, 10 proteins).

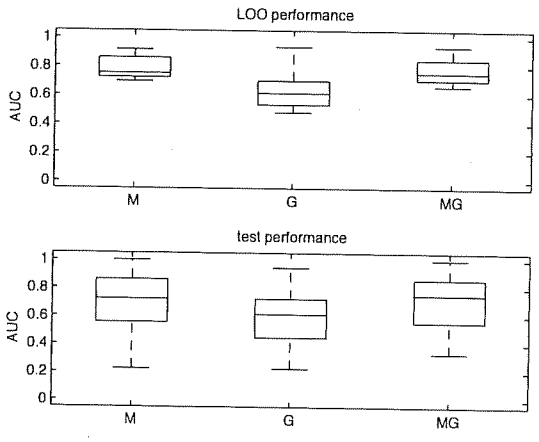


Figure 3.4: *Comparison for GRADE of the internal LOO performances (AUC) in subplot 1 and the corresponding test performances in subplot 2 using a boxplot representation across all 20 randomizations:* Boxplots are shown for the optimal number of selected features, ranging from 1 to 35 in steps of 5: M (30 genes), G (20 CNVs), and MG (35 genes, 1 CNV).

3.6.5 Correlation Analysis

We finally verified whether data from multiple layers of molecular biology are complementary by mapping the entities of the data sets based on their Entrez Gene IDs. Using the Spearman correlation coefficient, there was no significant correlation for rectal cancer between the abundance of the 90-92 proteins and their corresponding transcript at a significance level of 0.05, suggesting the importance of alternative splicing and post-translational modifications. For prostate cancer, up to three genes included in the models MG were significantly correlated with their corresponding CNV. The regions involving CNVs selected from the genomics data were also compared with the regions in which the selected genes from the microarray data were located. For those regions with the same rough chromosomal location, different genes were selected from both data sets.

3.7 Conclusion

In this chapter, we presented an integration approach in which experimental data from multiple levels in the genome are combined at the kernel level within the patient domain. To verify the merit of our integration approach over the use of a single -omics data set, models were built for the classification of cancer patients according to therapy response, prognostic factors, metastasis and recurrence. In

many studies, only a single data set is explored for the development of such profiles. However in our opinion, a single layer of molecular information is inadequate to explain the complete network of molecules underlying a disease. For both considered case studies on rectal and prostate cancer, the combination of high-throughput data sets (transcriptomics with proteomics, and transcriptomics with genomics) outperformed the models based on data from a single layer of biological information, independent of the outcome considered for prediction. For many of the selected genes, proteins and CNVs, involvement in rectal or prostate cancer was already defined, indicating the reliability of the obtained features. When more data sources will become available in the future, they can easily be added to our kernel-based integration framework.

By looking at correlation between data sets gathered from the same patients, we have shown that data from different layers are mainly complementary. Specific to rectal cancer, microarray data were not able to explain the complete molecular biology behind this cancer. When comparing models based on only microarray or proteomics data, the latter led to a better performance for certain outcomes and time points. This indicates that mRNA levels assessed by monitoring transcription with microarray technology do not always correspond to the levels or activity of the final gene products (that is, proteins) after translation.

We acknowledge that the models proposed in this chapter are quite expensive. There may exist other, less expensive data sets with complementary information as well. Firstly, clinical information such as tumor size, tumor location and number of positive lymph nodes is routinely gathered during therapy. However, we only had access to the clinical parameter age and verified whether this parameter could be of use. A univariate analysis based on the Wilcoxon rank sum test showed no significant difference in age according to the considered outcomes. In a multivariate logistic regression model, the parameter age was not significant either. Secondly, also imaging, such as computed tomography (CT) and magnetic resonance imaging (MRI), can be a potential predictor to be used in combination with high-throughput data sets. Our results emphasize the need for comprehensive multi-modal data gathered with high-throughput technologies and imaging, because it is unknown which levels of molecular biology are most relevant for prognostic prediction and should be combined in order to obtain affordable and clinically applicable models. We acknowledge that this will substantially increase costs in a first exploratory phase. However, this is a necessary investment to ultimately obtain cost-efficient models from which patients can directly benefit.

Chapter 4

New Kernel Function for Modeling Clinical Data

Despite the available high-throughput technologies, decisions concerning diagnosis, prognosis and therapy are for most examinations still based on clinical data such as age, gender and medical history. To improve clinical management, available patient information should be fully exploited. This requires appropriate modeling of relevant parameters. Concerning kernel methods, the linear kernel function is often used, despite its disadvantages as clinical data are a mix of variable types with each variable having its own range. We propose a new linear kernel function specifically adapted to the characteristics of clinical data, taking into account the type and range of each variable. Incorporated in a wLS-SVM, this clinical kernel function more accurately represents similarities between patients, resulting in significantly improved clinical decision support. This is illustrated by four clinical data sets within gynecology, as well as for the combination of clinical parameters and microarray data in three case studies on breast cancer.

4.1 Introduction

During a medical examination, patient-specific information such as age, menopausal status and medical history is registered. Histopathological parameters such as tumor size, lymph node status and relapse rate, and ultrasound data such as endometrium thickness are often registered as well, with the set of clinical

parameters characterizing a patient depending on the investigated disease. While clinical parameters are routinely measured by clinicians, high-throughput data are in general much more difficult and expensive to collect. A critical study on the prediction of breast cancer outcome has suggested that clinical markers and profiles obtained from high-throughput technologies have similar power for prognosis [107]. Moreover, clinical management is for many diseases and examinations fully guided by clinical data.

In this chapter, we therefore focus on an improved modeling of clinical data with kernel methods, to aid clinical decision support [77] [87]. In many previous clinical studies, the linear kernel function was used for this purpose due to its simplicity and interpretability [2] [221] [331]. Although favored by clinicians, several disadvantages occur when applied to clinical data. The influence of each variable on patients' similarity is proportional to its range, thereby undesirably enlarging the influence of irrelevant continuous variables and diminishing the contribution of important discrete variables. In Section 4.3.1, we therefore propose an alternative linear kernel function specifically developed for clinical data and more accurately representing similarities between patients. We show the improvement obtained with this alternative kernel function when applied to four clinical data sets within gynecology (Section 4.2.1 and 4.4). The advantage of our kernel function is furthermore illustrated when combining clinical and microarray data in three case studies on breast cancer (Section 4.2.2 and 4.4). As shown in Chapter 3, with the integration of heterogeneous data sets advantage can be taken from the strengths of the data. As in most microarray studies on cancer the focus is on the microarray analysis while available clinical data are not modeled in the same manner, we investigated in [79] the influence of our kernel function on data integration. Similar to Chapter 3, we opted for intermediate integration in which the data sets are treated as separate entities, instead of early integration where clinical variables would get lost within the huge amount of genes.

4.2 Data

In this section, an overview is given of four clinical data sets within gynecology, followed by a description of three microarray studies on breast cancer for which a sufficient amount of clinical data were available. For the clinical parameters, a distinction is made between continuous variables (labeled as C), discrete continuous variables with an intrinsic ordering and equal distance between two consecutive values (D), ordinal variables with an intrinsic ordering but lacking the equal distance between categories (O), and nominal variables without any ordering (N).

4.2.1 Clinical Data Sets

For the clinical data sets, several original variables were log-transformed or excluded in an unsupervised way. A logarithmic transformation reduces the asymmetry of a distribution so that an approximated normal distribution is obtained in which the influence of rarer patients with extreme values is reduced. Continuous variables with a positively skewed distribution were log-transformed. In case such a variable could take the value 0, the data were first augmented with parameter a before transformation. The choice of a should be based on the range of the variable so that the augmented values remain in the neighborhood of zero for the transformation to be effective. Due to the large range of the positively skewed variables in our data sets (such as tumor size and endometrium thickness), a was set to 1. For each data set, we also calculated the pairwise Spearman correlation coefficient between all variables. The pairs of variables with a correlation in absolute value above 0.7 were further investigated to reduce each data set to a set of variables lacking redundancy. Patients with incomplete data were excluded as imputation methods have only been defined for continuous variables.

Endometrial Disease

Data set I contains clinical information on 402 patients with an endometrial disease who underwent an ultrasound examination and color Doppler [334]. The patients were divided into two groups according to their histology: abnormal (hyperplasia, polyp, myoma and carcinoma) versus normal (proliferative endometrium, secretory endometrium, atrophia). The variable endometrium thickness obtained with a standard ultrasound and a color Doppler examination correlated with a coefficient of 0.97. As both variables were highly significant according to a univariate analysis (Wilcoxon rank sum test), we opted to include thickness obtained with the standard ultrasound examination. The variables parity (that is, number of times a woman has given birth) and gravidity (that is, number of times a woman has been pregnant) highly correlated as well ($\rho = 0.91$). Only parity was included, showing more relevance based on univariate results (p-value 0.32 vs. 0.63, respectively; Fisher's exact test). After excluding patients with incomplete data and correlation-based exclusion of redundant variables, data set I contained 22 variables for 339 patients, of which 163 were abnormal and 176 normal. An overview of the 22 clinical variables is given in Table 4.1.

Miscarriages

A prospective observational study of 1,828 women undergoing transvaginal ultrasound before 12 weeks gestation resulted in data for 2,356 pregnancies.

Among them 1,458 were normal at week 12 whereas 898 had miscarried by the end of week 12 [34]. The 18 clinical variables are shown in Table 4.2.

Pregnancies of Unknown Location

Data set III contains data on 1,003 pregnancies of unknown location (PUL) [139]. Within the PUL group, there are four clinical outcomes: failing PUL, intrauterine pregnancy (IUP), ectopic pregnancy (EP) and persisting PUL. As the last outcome is rare, persisting PULs (18 cases in the data set) were excluded, as well as pregnancies with missing data. The final data set consists of 856 PULs among which 460 failing PULs, 330 IUPs, and 66 EPs. Correct classification of EPs among PULs has been shown to be the most important diagnostic problem [66]. Correlation coefficients above 0.7 were observed between hCG (human chorionic gonadotropin) measured at 0 hrs and 48 hrs, progesterone measured at both time points and hCG ratio (48hrs/0hrs). Based on univariate results, the measurements at 48 hrs seemed most relevant for the problem at hand. Earlier measurements and hCG ratio were therefore excluded to avoid deteriorating performances due to the inclusion of redundant variables. The 12 remaining clinical variables are shown in Table 4.3.

Adnexal Masses

As fourth clinical data set, we studied a multicentric data set on adnexal masses, gathered by the IOTA group as introduced in Section 1.3 [324]. Before the start of the IOTA study, several scoring systems and mathematical models such as logistic regression models, artificial neural networks and support and relevant vector machines were developed to help less experienced ultrasonographers by distinguishing benign from malignant adnexal tumors [323]. A prospective validation on the data collected during IOTA study phase 1 revealed that the vector machine models performed best [211], with an area under the ROC curve exceeding 0.90 on an external data set [344]. During IOTA study phase 1, eleven new scoring systems and mathematical models were developed for calculating the risk of malignancy in adnexal tumors. These models were prospectively validated on internal data gathered during IOTA study phase 1b [343]. Also this validation study revealed that the SVM with linear kernel had the largest area under the ROC curve [331]. We wanted to verify whether this excellent performance could be improved by using the newly developed clinical kernel function.

Table 4.1: Clinical variables data set I (endometrial disease)

variable	type	range
1. age (years)	C	22 - 85
2. weight (kg)	C	45 - 160
3. number of miscarriages/abortions	D	0 - 5
4. parity	D	0 - 7
5. menopausal status	N	1,2,3
6. hormonal therapy	N	0,1,2,3,4
7. intrauterine device	N	0,1,2
8. type of AUB$^{\rho}$	N	1,2,3
9. amount of AUB$^{\rho}$	N	1,2,3
10. duration of AUB$^{\rho}$ (months)	C	0.406 - 4.80$^{\psi}$
11. endometrial cells	N	1,2,3
12. endometrium thickness on US$^{\epsilon}$ (mm)	C	0 - 3.91$^{\psi}$
13. intracavity fluid (mm)	C	0 - 8.7
14. 3-layer pattern	N	1,2
15. intracavity lesion	N	1,2,3
16. subendometrial cyst	N	1,2
17. endometrial cyst	N	1,2
18. number of calcifications	D	0 - 10
19. number of myoma	D	0 - 4
20. ovary aspect	N	1,2
21. presence of follicles	N	0,1
22. pedicle sign	N	1,2,3

$^{\rho}$AUB, abnormal uterine bleeding. $^{\epsilon}$US, ultrasound. $^{\psi}$On a logarithmic scale after correction for a positively skewed distribution ($log(x+1)$).

During phase 1 and 1b of the IOTA study, more than 40 clinical and ultrasound variables were collected from 1,573 patients, of whom 1,164 had a benign and 409 a malignant adnexal mass. We considered only those ultrasound variables that were included in at least one of the eleven previously developed models, resulting into 18 variables. Three of these variables were dummy variables describing the morphology of the lesion and could be replaced by morphology as such. A high correlation of 0.85 was observed between the maximal diameter of the lesion and the ovary. Due to lack of difference in univariate results, we opted for the maximal diameter of the lesion as this variable is included in the logistic regression model that will be prospectively validated in IOTA phase 3. An overview of the 15 remaining ultrasound variables is given in Table 4.4.

Table 4.2: Clinical variables data set II (miscarriages)

variable	type	range
1. age (years)	C	15 - 48
2. PBAC bleeding score	O	0 - 4
3. follow-up consent	N	0,1,2
4. ethnicity	N	0,1,2,3,4,5,6
5. regular dates	N	0,1,2
6. gravida	D	1 - 12
7. number of deliveries after 24 weeks	D	0 - 10
8. number of terminated pregnancies	D	0 - 4
9. number of early miscarriages	D	0 - 10
10. number of PULs$^{\rho}$	D	0 - 1
11. number of late miscarriages	D	0 - 5
12. number of ectopic pregnancies	D	0 - 1
13. previous chromosomal abnormalities	N	0,1
14. bleeding$^{\epsilon}$	N	0,1
15. pain$^{\epsilon}$	N	0,1
16. previous ectopic pregnancy$^{\epsilon}$	N	0,1
17. previous miscarriage$^{\epsilon}$	N	0,1
18. anxiety$^{\epsilon}$	N	0,1

$^{\rho}$PUL, pregnancy of unknown location. $^{\epsilon}$Indication for scan.

Table 4.3: Clinical variables data set III (pregnancies of unknown location)

variable	type	range
1. hCG$^{\rho}$ at 48h (U/l)	C	0 - 9.52$^{\psi}$
2. progesterone at 48h (nmol/l)	C	0 - 5.25$^{\psi}$
3. endometrium thickness (mm)	C	0.916 - 3.58$^{\epsilon}$
4. character of midline echo	N	0,1
5. free fluid in pouch of Douglas	N	0,1
6. gestational age (days)	C	2.30 - 4.61$^{\psi}$
7. lower abdominal pain	N	0,1
8. vaginal bleeding	N	0,1,2
9. previous miscarriage	N	0,1
10. previous ectopic pregnancy	N	0,1
11. anxiety	N	0,1
12. age (years)	C	14 - 49

$^{\rho}$hCG, human chorionic gonadotropin. $^{\psi}$On a logarithmic scale after correction for a positively skewed distribution ($log(x)$). $^{\epsilon}$On a logarithmic scale after correction for a positively skewed distribution ($log(x+1)$).

Table 4.4: Clinical variables data set IV (adnexal masses)

variable	type	range
1. age (years)	C	9 - 94
2. personal history of ovarian cancer	N	0,1
3. hormonal therapy	N	0,1
4. maximal diameter of the lesion (mm)	C	2.08 - 6.02$^{\psi}$
5. presumed ovarian origin of tumor	N	0,1
6. pelvic pain during examination	N	0,1
7. locularity of the tumor (morphology of the lesion)	N	0,1,2,3,4,5,6
8. maximal diameter of the solid component (mm)	C	0 - 50$^{\rho}$
9. number of papillary projections	D	0 - 4
10. blood flow within papillary projection	N	0,1
11. irregular internal cyst walls	N	0,1
12. acoustic shadows	N	0,1
13. color score of intratumoral blood flow	O	1 - 4
14. presence of venous blood flow only	N	0,1
15. presence of ascites	N	0,1

$^{\psi}$On a logarithmic scale after correction for a positively skewed distribution ($log(x)$).
$^{\rho}$maximal diameter of the solid component bounded to 50mm due to its binomial distribution (that is, the diameter equals 0 in those patients without a solid component).

4.2.2 Clinical and Microarray Data

Besides pure clinical data sets, we also considered three case studies in which both microarray data and a sufficient number of clinical variables were available. Breast cancer is one of the most extensively studied cancer types for which many microarray data sets are publicly available, with or without clinical information. Among them, we selected three data sets [56] [159] [310]. The microarray data were obtained with the Affymetrix technology and preprocessed with MAS5.0, the GeneChip Microarray Analysis Suite 5.0 software (Affymetrix). As probe selection for the Affymetrix gene chips relied on earlier genome and transcriptome annotation that are significantly different from current knowledge, an updated array annotation was used for the conversion of probes to Entrez Gene IDs, lowering the number of false positives [89] [213]. Finally, the low signal-to-noise ratio of microarray data was taken into account by unsupervised exclusion of genes with low variation, retaining the 5,000 most varying genes.

The first data set of 129 patients contained information on 17 clinical variables [56]. Five variables were excluded: two redundant variables that were least informative based on univariate analyses in those variable pairs with a correlation coefficient exceeding 0.7, and three variables with too many missing values. After exclusion

of patients with missing clinical information, this data set (referred to as data set V and represented in Table 4.5) consisted of 110 patients, in 85 of whom disease didn't recur whilst in 25 patients disease recurred.

The second data set in which response to treatment was studied, consisted of 12 variables for 133 patients [159]. Exclusion of 3 variables with too many missing values, 1 redundant variable according to correlation, and 4 patients lacking information for some of the 8 remaining variables resulted in data set VI. Of the 129 remaining patients, 33 showed complete response to treatment while 96 patients were characterized as having residual disease. An overview of the variables is provided in Table 4.6.

In the last case study, relapse was studied in 187 patients [310]. One variable needed to be excluded due to too many missing values, while information on the remaining variables was incomplete for 10 patients. Data set VII therefore contained information on 5 variables for 177 patients. In 112 patients, no relapse occurred while 65 patients were characterized as having a relapse. We refer to Table 4.7 for an overview of the variables included in this case study.

Table 4.5: Clinical variables data set V (breast cancer - recurrence)

variable	type	range	p-value
1. age (years)	C	31 - 88	0.5472*
2. ethnicity	N	0,1,2	0.1507°
3. ER$^{\rho}$ status	N	0,1	0.049°
4. PR$^{\epsilon}$ status	N	0,1	0.1996°
5. radiation treatment	N	0,1	0.0927°
6. chemotherapy	N	0,1	0.5326°
7. hormonal therapy	N	0,1	0.6741°
8. nodal status (N)	O	0 - 2	0.0001*
9. metastasis (M)	N	0,1	3.6e-4°
10. tumor stage	O	1 - 4	0.0002*
11. tumor size (cm)	C	0.262 - 2.14$^{\psi}$	<0.0001*
12. tumor grade	O	1 - 3	0.0601*

$^{\rho}$Estrogen receptor. $^{\epsilon}$Progesterone receptor. $^{\psi}$On a logarithmic scale after correction for a positively skewed distribution ($log(x+1)$). *Wilcoxon rank sum test. °Fisher's exact test.

Table 4.6: Clinical variables data set VI (breast cancer - treatment response)

variable	type	range	p-value
1. age (years)	C	28 - 79	0.0448*
2. ethnicity	N	0,1,2,3,4	0.5791°
3. pretreatment tumor stage	O	1 - 4	0.7424*
4. nodal status (N)	O	0 - 3	0.7634*
5. nuclear grade	O	1 - 3	0.0005*
6. ER$^{\rho}$ status	N	0,1	2.0e-8°
7. PR$^{\epsilon}$ status	N	0,1	6.2e-4°
8. HER2$^{\psi}$ status	N	0,1	0.0368°

$^{\rho}$Estrogen receptor. $^{\epsilon}$Progesterone receptor. $^{\psi}$Human epidermal growth factor receptor 2. *Wilcoxon rank sum test. °Fisher's exact test.

Table 4.7: Clinical variables data set VII (breast cancer - relapse)

variable	type	range	p-value
1. age (years)	C	32 - 86	0.3307*
2. tumor size (cm)	C	0 - 2.22$^{\psi}$	0.0015*
3. nodal status	N	0,1	0.3007°
4. ER$^{\rho}$ status	N	0,1	0.0310°
5. tamoxifen treatment	N	0,1	0.1970°

$^{\rho}$Estrogen receptor. $^{\psi}$On a logarithmic scale after correction for a positively skewed distribution ($log(x+1)$). *Wilcoxon rank sum test. °Fisher's exact test.

4.3 Methodology

After having described the data sets, we now present a new kernel function, specifically adapted to the characteristics of clinical data.

4.3.1 Kernel Function for Clinical Data

A kernel function normalized according to (2.3) provides a measure of similarity between patients based on their clinical profiles. The obtained similarity value with the normalized linear kernel function, however, strongly depends on the range of each variable, favoring continuous variables with a large range (for example, age from 20 to 50 years contrary to progesterone from 0 to 5 nmol/l). Also for ordinal variables the comparison of two patients with value 1 and 2 depends on the range of this variable. These patients will be less similar when the variable has only three categories compared to six. Furthermore, when an ordinal variable equals

zero, the inner product will always be zero, independent of patients' dissimilarity. For nominal variables that lack an intrinsic ordering, the inner product between two patients should only be larger than zero in case both patients have the same category.

The scale of the input data was already known to influence model performance. A rough distinction according to variable type was incorporated in the LS-SVMlab toolbox [246]. Binary variables were re-scaled to $\{-1, 1\}$ whilst continuous variables were normalized, avoiding attributes in larger numeric ranges to dominate those in smaller ranges. Other variables, however, were kept unchanged, thereby not distinguishing ordinal from nominal variables. Although for the linear kernel function an improvement can be expected by standardizing the continuous variables (in the tables and figures referred to as *linear std*), we introduce an alternative linear kernel function, specifically developed for clinical data and referred to as the *clinical kernel function*. A distinction is made between (discrete) continuous, ordinal and nominal variables, and per variable type a kernel function is defined. To guarantee the same influence of each variable, the appropriate kernel function is applied to each variable individually before calculating the global, heterogeneous kernel matrix. The following notations are used: $k_z(i,j)$ denotes the kernel function for variable z between patients i and j; $K_z(i,j)$ $\forall i,j$ represents the corresponding individual kernel matrix for variable z; and $K(i,j)$ $\forall i,j$ represents the global, heterogeneous kernel matrix. For clinical studies in which the data are nonlinear separable, the polynomial kernel $({x^i}^T x^j + \tau)^d$ can easily be adapted in the same way by replacing ${x^i}^T x^j$ by the clinical kernel function [330]. For both the linear and polynomial kernel function, the more robust normalized version (2.3) was used, although not explicitly mentioned in the remaining of this chapter.

(Discrete) Continuous and Ordinal Clinical Variables

The ordinal variables in the considered data sets are bleeding score, color score, tumor stage, tumor grade and nodal status. For those variables, the categories were replaced by their rank. Under the assumption of an equal distance between 2 consecutive categories, the same kernel function is proposed for (discrete) continuous and ordinal variables.

$$k_z(i,j) = \frac{r - |z_i - z_j|}{r}, \tag{4.1}$$

with constant r the range or the number of categories for variable z, extracted from the data or based on clinical knowledge or *a priori* information from specialist literature. The difference in z-value between two patients i and j is compared with and rescaled to range r.

Nominal Clinical Variables

For nominal variables, the kernel function between patients i and j is defined as the Kronecker delta function

$$k_z(i,j) = \begin{cases} 1 & \text{if } z_i = z_j \\ 0 & \text{if } z_i \neq z_j \end{cases} . \tag{4.2}$$

This kernel function is independent of the variable values, making binary dummy variables obsolete.

Final Kernel for Clinical Data

For each individual kernel function k_z, the similarity measure is forced to the interval [0,1]. The global, heterogeneous kernel matrix K can therefore be defined as the sum of the individual kernel matrices K_z divided by the total number of clinical variables p, and it describes the similarity for a class of patients based on a set of variables of different types. This corresponds to the additive kernel function $k(i,j) = \frac{1}{p}\sum_{z=1}^{p} k_z(i,j)$.

Proof Validness Clinical Kernel

For the clinical kernel function to be general applicable, it needs to comply with the following kernel properties [136]: symmetry (that is, $k(x^i,x^j) = k(x^j,x^i)$ for any two objects x^i, x^j), Cauchy-Schwartz inequality ($k^2(x^i,x^j) \leq k(x^i,x^i)k(x^j,x^j)$), and positive semidefiniteness ($s^T K s \geq 0$ for all non-zero vectors $s \in \mathbb{R}^N$). The kernel variant for nominal variables (4.2) is defined as the Kronecker delta function, which is a valid kernel [136]. The kernel variant for continuous and ordinal variables (4.1) is a kernel function because it can be derived from the triangular kernel based on the following two properties [71]:

Property 4.1 (Weighted sum) *If k_1 and k_2 are two positive definite kernels with a_1 and a_2 positive real numbers, then $k(x^i,x^j) = a_1 k_1(x^i,x^j) + a_2 k_2(x^i,x^j)$ is a positive definite kernel.*

Property 4.2 ($\mathbb{R}^p$-valued function) *If k_3 is a positive definite kernel and g an $\mathbb{R}^p$-valued function on $X \in \mathbb{R}^N$, then $k(x^i,x^j) = k_3(g(x^i),g(x^j))$ is a positive definite kernel.*

The triangular kernel function is defined as $k(u) = 1 - |u|$ for $|u| \leq 1$ and 0 otherwise, represented as $k(u) = (1-|u|)1_{(|u|\leq 1)}$. By replacing u with the

difference in z-value between patients i and j, $k(u)$ becomes $k_z(i,j) = (1 - |z_i - z_j|)1_{(|z_i - z_j| \leq 1)}$. Applying Property 4.2 with $g = 1/r$ results in

$$k_z(i,j) = \left(1 - \left|\frac{z_i}{r} - \frac{z_j}{r}\right|\right) 1_{(|z_i - z_j| \leq r)} = \left(\frac{r - |z_i - z_j|}{r}\right) 1_{(|z_i - z_j| \leq r)}.$$

Finally, the global kernel is a valid kernel function due to Property 4.1.

Example

We first calculate the similarity (that is, kernel matrix) between three patients h, i, and j for the continuous variable age. Patient h is 23 years old, patient i 26, and patient j 54. Based on the data set, the minimal age is 20 and maximal age 100. The elements in the kernel matrix can then be calculated as follows:

$$k_{age}(h,i) = (80 - |23 - 26|)/80 = 77/80$$

$$k_{age}(h,j) = (80 - |23 - 54|)/80 = 49/80$$

$$k_{age}(i,j) = (80 - |26 - 54|)/80 = 52/80$$

The resulting kernel matrix for variable age equals

$$K_{age} = \begin{bmatrix} 1 & 0.9625 & 0.6125 \\ 0.9625 & 1 & 0.6500 \\ 0.6125 & 0.6500 & 1 \end{bmatrix}$$

with values decreasing with increasing dissimilarity between patients.

The extent of most types of cancer is described with a TNM classification system: T represents the size of the primary tumor with ranks 1, 2, 3 and 4; N describes the degree of spread of the tumor to regional lymph nodes (0, 1 or 2); the absence or presence of metastasis to distant organs beyond the regional lymph nodes is represented by the binary variable M. Patient h is characterized by T1N0M0, patient i by T3N2M1 and patient j by T4N1M1. The resulting individual kernel matrices are

$$K_T = \begin{bmatrix} 1 & 0.33 & 0 \\ 0.33 & 1 & 0.66 \\ 0 & 0.66 & 1 \end{bmatrix}, \; K_N = \begin{bmatrix} 1 & 0 & 0.5 \\ 0 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{bmatrix},$$

$$K_M = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix}.$$

The proposed kernel function takes into account the range of variables (for example, 0.66 (T) vs. 0.5 (N) for a difference of one unit in the range). Moreover, the kernel value equals zero when two patients are most dissimilar (T1 vs. T4, and N0 vs. N2). The linear kernel function, on the other hand, would have led to erroneous positive values.

The heterogeneous kernel matrix for the similarity between patients h, i and j based on age, tumor size (T), lymph node spread (N), and metastasis (M) is given by

$$K = \frac{1}{4}(K_{age} + K_T + K_N + K_M) = \begin{bmatrix} 1 & 0.324 & 0.2781 \\ 0.324 & 1 & 0.7042 \\ 0.2781 & 0.7042 & 1 \end{bmatrix}.$$

4.3.2 Model Building

For the clinical data sets (Section 4.2.1), the data were randomly split into $2/3^{rd}$ for training and $1/3^{rd}$ for testing with stratification to outcome. On the training samples, a 10-fold CV approach with grid search was applied for the optimization of the regularization parameter γ. When the polynomial kernel function was used, a three-dimensional grid was required for the additional optimization of tuning parameters τ and d, both varying on a linear scale from 1 to 5. Contrary to the typical use of the polynomial kernel with $\tau = 1$ [40] [128] [371], scaling (that is, $\tau \neq 1$) was considered as this has shown to increase test performance [339]. The optimal model with the highest 10-fold train AUC was further validated on the $1/3^{rd}$ of samples left out for testing. To obtain a better estimate of the prediction performance of the classifiers and thus a more reliable comparison of the clinical alternative with the traditional kernel functions, the split of the data in a training and test part was repeated 100 times. The 10-fold train AUC and test AUC values averaged over these 100 repetitions are reported. The one-sided paired-sampled t-test was used to compare the AUC values obtained with the applied kernel functions. A p-value of 0.05 was considered statistically significant.

For the three clinical data sets for which corresponding microarray data are available (Section 4.2.2), the number of available samples is much smaller (in the order of hundreds as compared to thousands of samples in pure clinical studies). The intercept constant τ in the polynomial kernel function was therefore fixed to 1. Moreover, instead of selecting an independent test set, 10-fold CV was applied to the full data set and repeated 100 times. For each repetition, the random division

of data into 10 folds was performed with stratification to outcome. For the kernel matrix obtained from microarray data, the 200 most differentially expressed genes selected from the training data with the Wilcoxon rank sum test were considered. The performance of the clinical kernel with respect to the linear kernel function was independent of the specific number of incorporated genes.

In the case studies with clinical and microarray data, three settings were considered for the evaluation of the clinical kernel function: only the clinical data sets were considered for classification (1 CL + 0 MA), the influence of microarray and clinical data on prediction was set equal (1/2 CL + 1/2 MA), and the weights assigned to both data sets were optimized (μ CL + $(1-\mu)$ MA). In the latter, we made μ vary from 0 to 1 in steps of 0.05, and the combination (γ,μ) or (γ,μ,d) that led to the largest 10-fold AUC was selected. The weights assigned to the matrices reflect the importance of each individual data set for the problem at hand as the normalized linear and polynomial kernel functions were used and the kernel values obtained with the clinical alternative lay between 0 and 1 due to their construction.

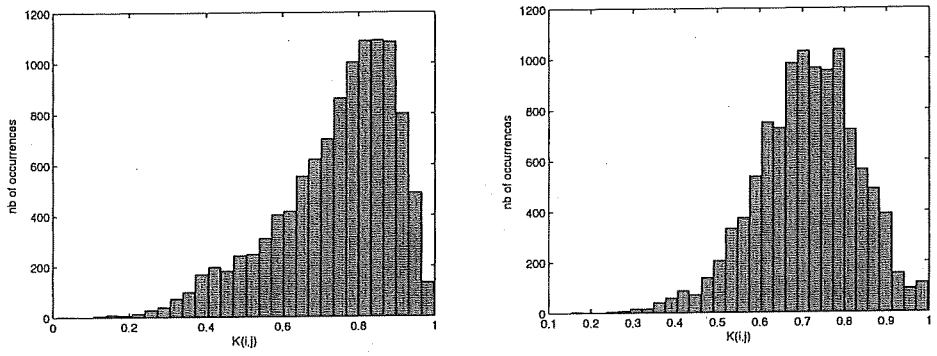
4.4 Results

In this section, the linear and clinical kernel function are first studied in detail. The results are then presented for the clinical studies and for the combination of clinical and microarray data. We end this section by showing the advantage of improved modeling in a broader spectrum.

4.4.1 Comparison of the Linear and Clinical Kernel Function

In a first phase, we verified whether the clinical kernel function better represents true similarity between patients. For this purpose, a publicly available data set on breast cancer containing a mix of continuous, ordinal and nominal variables was used in which the appearance of distant subclinical metastases was predicted based on the primary tumor [332]. This data set of 148 patients contained 13 clinical parameters, represented in Table 4.8 [54]: 2 continuous parameters, age (20-60 years old) and tumor diameter (0-70 mm); 1 discrete continuous parameter, ranging from 0 to 15; 3 ordinal parameters with 3 categories each; and 7 nominal, binary parameters.

Figure 4.1 shows the histograms of the kernel matrices based on the linear and clinical kernel function. The histogram for the linear kernel function after standardization of the continuous variables has a larger median equal to 0.7766 compared to a median of 0.7168 for the clinical kernel function. Four comparisons based on the patients' data shown in Table 4.8 were made to verify whether differences in the kernel values correspond to true differences in patient data.



(a) Linear kernel (b) Clinical kernel

Figure 4.1: *Histograms of the kernel matrices:* **(a)** based on the linear kernel function after standardization of the continuous variables, **(b)** based on the clinical kernel function.

Table 4.8: Specific patient data from [332]

	Comparison	a		b		c			d			
	Patient	55	84	36	55	1	18	45	6	58	3	79
C_1	age (years)	50	41	37	50	41	41	41	40	39	48	49
C_2	tumor size (mm)	8	45	50	8	20	20	20	14	15	15	16
D_1	nb pos lymph nodes	0	4	0	0	0	0	1	0	0	0	0
O_1	N (pN0, 1-3, $\geq$4)	1	3	1	1	1	1	2	1	1	1	1
O_2	grade	3	1	1	3	1	1	1	1	1	1	1
O_3	NIH risk	1	3	3	1	3	3	3	3	3	3	3
N_1	mastectomy	0	1	1	0	0	1	0	0	0	0	0
N_2	estrogen receptor	1	0	0	1	1	0	1	1	1	1	1
N_3	chemotherapy	0	1	0	0	0	0	0	0	0	0	0
N_4	hormonal therapy	0	1	0	0	0	0	0	0	0	0	0
N_5	St. Gallen criteria	0	1	1	0	1	1	1	1	1	1	1
N_6	NIH consensus	0	1	1	0	1	1	1	1	1	1	1
N_7	T ($\leq$2cm or >2cm)	0	1	1	0	0	0	0	0	0	0	0

Comparison a) Minimal Clinical Kernel Value

Patients 55 and 84 are most dissimilar according to the clinical kernel function with a kernel value of 0.1523 contrary to 0.1415 for the linear kernel function (see Table 4.8 for the clinical parameters). These two patients not only differ greatly in tumor size (C_2), they are also most different for the ordinal variables and are distinct concerning all nominal variables.

Comparison b) Minimal Linear Kernel Value

The lowest value for the linear kernel function, on the other hand, occurs between patients 36 and 55 and is equal to 0.0089, in contrast to a clinical kernel value of 0.3904. When comparing the data of these two patients, age and tumor size seem to be more different than in comparison a. These patients, however, are equal regarding the discrete continuous variable, one ordinal and two nominal variables. The clinical kernel function ranks these patients as more similar due to the equal weight given to each variable. In the linear kernel, however, the influence of the continuous variables age and tumor size dominates the influence of the non-continuous variables.

Comparison c) Influence of Ordinal and Nominal Variables

To validate the influence of non-continuous variables on the kernel matrix, we chose three patients with the same age and tumor size (see Table 4.8). Patients 1 and 18 are different according to two nominal variables, whilst patients 1 and 45 slightly differ in the discrete continuous variable (0 vs. 1 on a range from 0 to 15) and one ordinal variable (1 vs. 2 for O_1 with 3 categories). Taking into account the range of the variables, patients 1 and 45 are more similar than patients 1 and 18. This difference in similarity is much clearer with the clinical kernel function (0.9564 and 0.8462, respectively) than with the linear kernel function (0.9465 and 0.9313, respectively).

Comparison d) Influence of Continuous Variables

Finally, we compared four patients with the same ordinal and nominal variables but some minor differences in age and tumor size. Patients 6 and 58 differ 1 year in age with a difference of 1mm in tumor size. The same differences hold for patients 3 and 79, but they are both older with a slightly larger tumor. For the clinical kernel function, the similarities $k(6, 58)$ and $k(3, 79)$ are both equal to 0.9970. These similarities, however, are slightly different according to the linear kernel function (0.9983 and 0.9984).

Concluding remark

These comparisons of the linear and clinical kernel function show that differences in kernel values obtained with the linear kernel function do not optimally reflect true differences in patient data. This is caused by continuous variables dominating non-continuous ones and because the range of the variables is not taken into account. This results in assigning patients to be similar when only ordinal and nominal variables differ, or dissimilar when more differing with respect to the continuous variables. In general, we can conclude from these comparisons that the clinical kernel provides a better representation of patients' similarity by giving an equal weight to all variables and taking into account the range of the variables.

4.4.2 Results for the Clinical Data Sets

We compared the linear and polynomial kernel function with the clinical alternatives on four data sets when used in a supervised classification algorithm. A 10-fold CV approach was applied for training a wLS-SVM model, subsequently validated on a test set. The 10-fold train and test results averaged over 100 random repetitions are shown in Table 4.9 while the corresponding boxplots are provided in Figure 4.2. For all data sets, the wLS-SVM models based on the clinical kernel definition significantly outperformed the models based on the linear and polynomial kernel (with and without standardization of the continuous variables). When comparing the polynomial kernel with its clinical variant, the intercept and degree were - in the majority of repetitions - 1 for the polynomial kernel, while varying from 1 to 5 and 3 to 5, respectively for the nonlinear clinical kernel. When comparing the clinical alternatives, in three out of four case studies the nonlinear clinical kernel outperformed the linear version on the training data; however, it produced worse test results for all four data sets. The linear clinical kernel function has thus a better generalization performance. For the linear clinical kernel, the increase in 10-fold train AUC values ranged from -0.01 to 0.29, whereas the interval was [-0.06; 0.35] for the test AUC values. For the nonlinear version, the intervals were [-0.01; 0.28] and [-0.15; 0.24], respectively.

Table 4.9: Results for four clinical data sets within gynecology

Data set	Kernel function	10-fold AUC (std)	p-value°	test AUC (std)	p-value°
I					
	linear	0.7423 (0.0226)	1.1e-46	0.7501 (0.0372)	3.3e-24
	linear std$^{\rho}$	0.7391 (0.0240)	1.9e-43	0.7508 (0.0370)	1.1e-19

Continued on next page

Table 4.9 – continued from previous page

Data set	Kernel function	10-fold AUC (std)	p-value°	test AUC (std)	p-value°
	clinical	0.7865 (0.0230)		0.7911 (0.0381)	
	poly	0.7314 (0.0251)	2.3e-47	0.7376 (0.0386)	2.8e-20
	poly std$^{\rho}$	0.7309 (0.0244)	3.9e-49	0.7379 (0.0385)	1.3e-19
	clin poly	0.7798 (0.0228)		0.7798 (0.0431)	
II					
	linear	0.7519 (0.0084)	3.8e-66	0.7542 (0.0140)	1.1e-43
	linear std$^{\rho}$	0.7320 (0.0073)	6.5e-80	0.7343 (0.0160)	1.8e-54
	clinical	0.7771 (0.0082)		0.7779 (0.0133)	
	poly	0.7354 (0.0088)	2.6e-99	0.7369 (0.0143)	4.7e-55
	poly std$^{\rho}$	0.7514 (0.0094)	9.8e-89	0.7539 (0.0155)	9.3e-33
	clin poly	0.8201 (0.0085)		0.7726 (0.0151)	
III					
	linear	0.6769 (0.0282)	1.6e-73	0.6880 (0.0521)	7.3e-43
	linear std$^{\rho}$	0.6234 (0.0420)	7.8e-69	0.6185 (0.0585)	1.2e-54
	clinical	0.8187 (0.0221)		0.8147 (0.0376)	
	poly	0.6477 (0.0303)	1.9e-60	0.6581 (0.0582)	3.6e-24
	poly std$^{\rho}$	0.6810 (0.0371)	1.3e-66	0.7043 (0.0626)	2.0e-9
	clin poly	0.8343 (0.0217)		0.7543 (0.0715)	
IV					
	linear	0.9116 (0.0059)	2.4e-98	0.9109 (0.0120)	1.5e-72
	linear std$^{\rho}$	0.9221 (0.0052)	7.3e-92	0.9219 (0.0107)	1.0e-61
	clinical	0.9453 (0.0044)		0.9439 (0.0094)	
	poly	0.9041 (0.0063)	1.2e-100	0.9040 (0.0119)	8.8e-67
	poly std$^{\rho}$	0.9226 (0.0049)	2.0e-94	0.9225 (0.0107)	8.3e-38
	clin poly	0.9480 (0.0042)		0.9362 (0.0106)	

°One-sided paired-sampled t-test for the comparison of the linear and polynomial kernel with the clinical alternative. $^{\rho}$Continuous variables were normalized before applying the kernel function.

4.4.3 Results for Data Integration

The results for the three case studies with clinical and microarray data are shown in Table 4.10 and Figure 4.3. When only clinical data are considered, the same trend as with the 4 previous data sets was observed, that is, a significant increase in performance was obtained with the clinical kernel definition, except for data set VI for which the polynomial kernel function performed better. When combined with microarray data, the clinical kernel variant also resulted in a significant

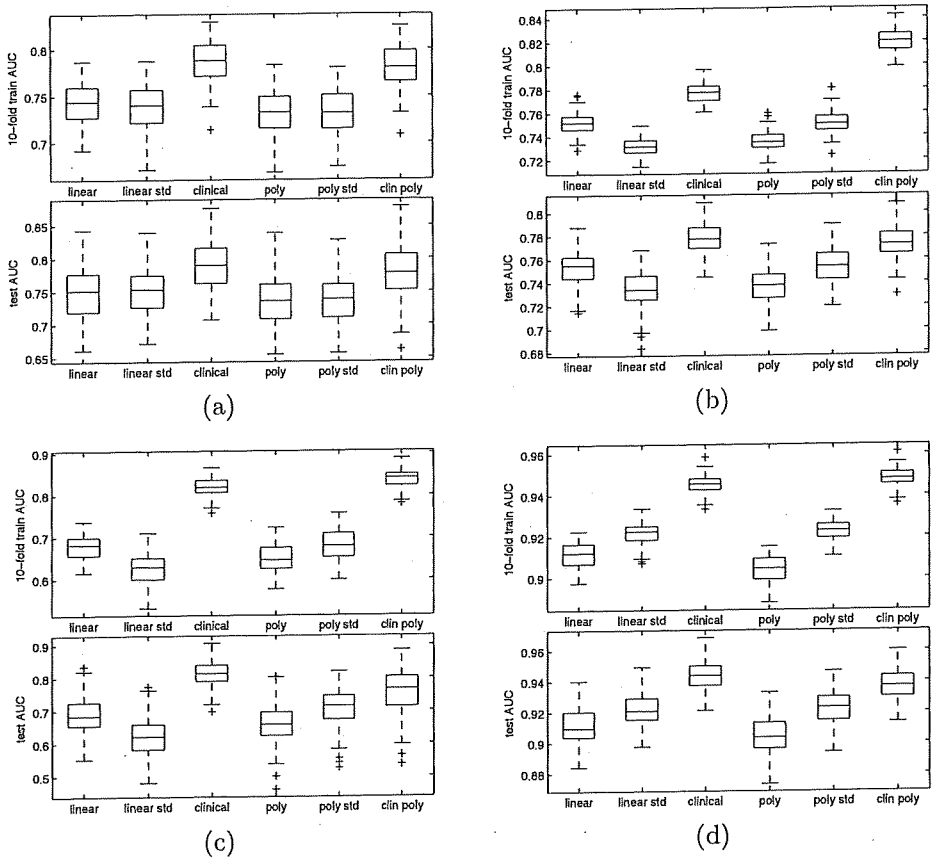


Figure 4.2: *Boxplots of the 10-fold train and test AUC values obtained in 100 repetitions for 4 clinical data sets on (a) endometrial disease, (b) miscarriages, (c) pregnancies of unknown location, and (d) adnexal masses.*

improvement, both for an equal influence of clinical and microarray data and with the weights assigned to both data sets optimized. Overall, the clinical variant of the polynomial kernel performed better than the linear clinical kernel. For data set V and VII, however, a degree of 1 already led to optimal results, both for μ optimized and set to 1. Results are shown when the 200 most differential genes were considered for the calculation of the microarray-based kernel matrix. A similar trend was observed with the inclusion of less (20, 50, 100) and more (500) genes.

Figure 4.4 shows per case study the histogram of the weights assigned to the clinical data set when combined with microarray data for the linear and clinical kernel function. Similar distributions were obtained for the polynomial kernel

function and its clinical variant. Tables 4.5, 4.6 and 4.7 contain the univariate results for the clinical parameters. A clear link was observed between the weights and the significance of the continuous variables (age, tumor size). With the linear kernel function, the influence of continuous variables with a wide range is much larger on the calculated patients' similarities than that of ordinal and nominal variables. When these continuous variables are in addition significantly related to the predicted outcome, the corresponding kernel matrix is assigned a large weight. This was the case for data set V caused by the relevance of tumor size. However, due to the equal contribution of variables in the clinical kernel function, the corresponding kernel matrix contains informative patients' similarity values in which the relevance of all types of variables is taken into account (see also Figure 4.1). Thus, although in one case study a lower weight was given to the kernel matrix based on the clinical kernel function, it performed better in combination with microarray data because on top of significant continuous variables, other ordinal and nominal variables were also relevant for the problems at hand.

4.4.4 Importance of the Clinical Kernel in a Broader Spectrum

We additionally verified whether the clinical kernel function outperforms the linear and polynomial kernel when using the SVM classifier [378]. The regular SVM was applied to data sets I and II, while the weighted SVM was implemented for the highly unbalanced data set III. As shown in [378], the clinical kernel outperformed the linear and homogeneous polynomial kernel with degree 2, 3 and 4 for data set I and II, while the linear kernel function in combination with weighted SVM led to a higher AUC for data set III.

In [378], we also focused on multiple kernel learning (MKL) for the combination of multiple data sets or multiple kernel matrices derived from the same data using distinct kernel functions. For data sets I to III, eight kernel matrices were created based on the linear kernel, clinical kernel, three RBF kernels and three homogeneous polynomial kernels. The latter two kernel types varied in width and degree, respectively. Among eight MKL algorithms that were exploited, the approach in which the weights assigned to the different kernel matrices were optimized and smoothly distributed over the matrices, was optimal. For all data sets, the optimized weights showed a pattern related to the performance of the individual kernels, with dominant weights assigned to the best kernels. The clinical kernel function was assigned the largest weight, except for data set II for which one of the RBF kernels was more informative.

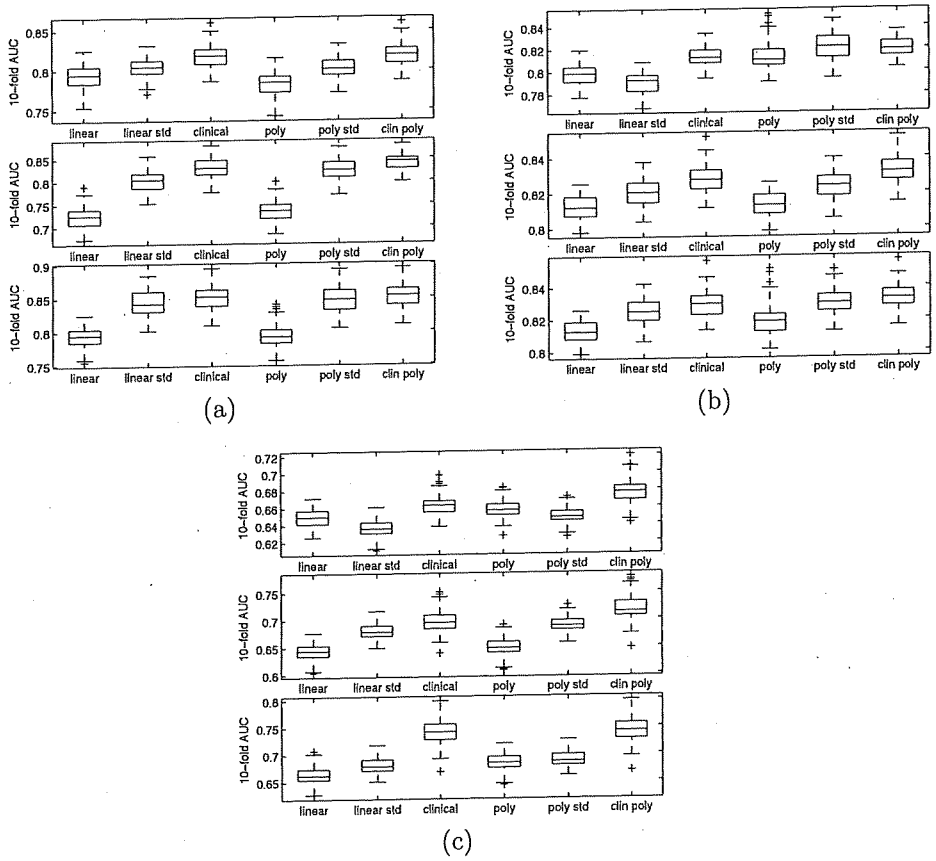


Figure 4.3: *Boxplots of the 10-fold AUC values obtained in 100 repetitions for three case studies on breast cancer, for the prediction of (a) recurrence, (b) treatment response, and (c) relapse:* For each case study, results are shown when only clinical data are considered (1 CL +0 MA; top), when clinical and microarray data have the same influence on prediction (1/2 CL +1/2 MA; middle), and when the weights assigned to both data sets are optimized (μ CL $+(1-\mu)$ MA; bottom).

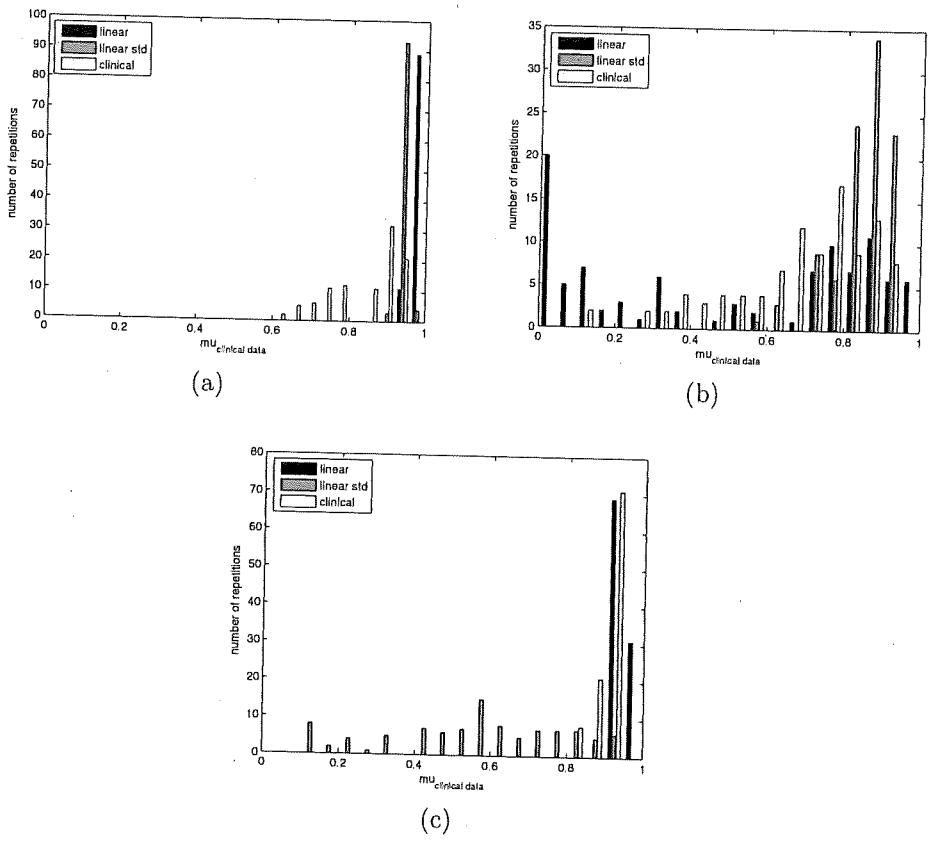


Figure 4.4: *Histogram of the weights assigned to the clinical data in 100 repetitions when combined with microarray data for the prediction of (a) recurrence, (b) treatment response, and (c) relapse:* Weights are shown when assigned to the kernel matrix obtained with the linear kernel function (black), with the linear kernel function after standardization of the continuous variables (dark gray), and with the clinical kernel function (light gray).

Table 4.10: Results on data integration for three case studies on breast cancer

Data	Kernel	1 CL + 0 MA		1/2 CL +1/2 MA		μ CL +$(1-\mu)$ MA	
set	function	10-fold AUC (std)	p-value°	10-fold AUC (std)	p-value°	10-fold AUC (std)	p-value°
V							
	linear	0.7932 (0.0151)	9.2e-34	0.7239 (0.0248)	1.1e-83	0.7941 (0.0133)	2.6e-57
	linear std$^{\rho}$	0.8045 (0.0119)	1.5e-18	0.8041 (0.0232)	2.4e-46	0.8446 (0.0192)	6.5e-07
	clinical	0.8180 (0.0150)		0.8317 (0.0212)		0.8507 (0.0168)	
	poly	0.7823 (0.0143)	1.6e-52	0.7348 (0.0245)	3.9e-77	0.7935 (0.0167)	4.3e-67
	poly std$^{\rho}$	0.8018 (0.0128)	3.9e-25	0.8237 (0.0236)	1.9e-19	0.8458 (0.0195)	1.3e-5
	clin poly	0.8180 (0.0150)		0.8402 (0.0179)		0.8508 (0.0168)	
VI							
	linear	0.7988 (0.0088)	9.9e-39	0.8127 (0.0071)	1.1e-47	0.8133 (0.0069)	5.5e-41
	linear std$^{\rho}$	0.7903 (0.0099)	6.3e-53	0.8203 (0.0078)	1.9e-32	0.8253 (0.0075)	4.9e-11
	clinical	0.8132 (0.0083)		0.8278 (0.0081)		0.8289 (0.0085)	
	poly	0.8122 (0.0117)	1.1e-10	0.8131 (0.0071)	7.2e-48	0.8183 (0.0088)	4.7e-32
	poly std$^{\rho}$	0.8206 (0.0113)	0.936	0.8236 (0.0074)	7.3e-25	0.8299 (0.0073)	3.7e-7
	clin poly	0.8193 (0.0077)		0.8317 (0.0072)		0.8326 (0.0075)	
VII							
	linear	0.6497 (0.0099)	3.0e-18	0.6431 (0.0152)	1.0e-60	0.6651 (0.0172)	1.7e-66
	linear std$^{\rho}$	0.6363 (0.0092)	1.9e-40	0.6796 (0.0135)	1.9e-21	0.6803 (0.0136)	9.1e-53
	clinical	0.6624 (0.0116)		0.6971 (0.0197)		0.7418 (0.0229)	
	poly	0.6574 (0.0099)	9.1e-27	0.6487 (0.0171)	2.0e-63	0.6851 (0.0172)	4.6e-54
	poly std$^{\rho}$	0.6486 (0.0084)	2.9e-38	0.6880 (0.0133)	4.0e-30	0.6890 (0.0132)	1.5e-47
	clin poly	0.6765 (0.0141)		0.7179 (0.0222)		0.7420 (0.0228)	

°One-sided paired-sampled t-test for the comparison of the linear and polynomial kernel with the clinical alternative.

$^{\rho}$Continuous variables were normalized before applying the kernel function.

4.5 Conclusion

When applying the normalized linear or polynomial kernel function for modeling clinical data, the influence of each variable is proportional to its range. When the variables with the largest range are most informative for the target outcome, good results may be obtained. However, when the reverse is true, this kernel function is not optimal. Correlation with outcome is often unknown beforehand, nominal variables with numerous categories can distort the calculation of patient similarity, and moreover, dependency on variable range should be discarded. Hence, each variable should have the same influence on the calculation of patient similarities, which was previously not the case. We therefore proposed a linear additive kernel function as alternative for the linear and polynomial kernel function that takes into account the type and range of each variable. This requires the specification of each type of variable, as well as the range for continuous variables and the number of categories for ordinal variables based on the data or *a priori* knowledge.

From our results, we can conclude that the clinical kernel function represents similarities between patients more accurately. Moreover, the wLS-SVM based on the clinical kernel variant significantly outperformed the linear and polynomial kernel function when tested on four pure clinical data sets and three sets of clinical parameters gathered in microarray studies. When in the latter case studies microarray data were added by using our kernel-based integration approach presented in Chapter 3, the clinical kernel variant also led to a significant increase in performance. Finally, the kernel function proposed in this chapter is not limited to clinical data. Any data set consisting of different types of variables can benefit from this function.

Chapter 5

Modeling Array CGH Data with HMM-Based Feature Selection

Little attention has been paid to genome-based classification. However, to be able to incorporate array CGH data in our kernel-based integration framework, chromosomal regions characterizing a group of patients were gathered using recurrent hidden Markov Models (HMM) and reduced to a relevant subset of features for classification. For three case studies on ovarian cancer, lung cancer and oral squamous cell carcinoma, wLS-SVM models resulted in LOO or 10-fold CV accuracies for the prediction of cancer subtype, ranging from 88% to 95.5%. The combination of recurrent HMMs for the detection of copy number alterations with wLS-SVM classifiers offers a novel methodological approach for classification based on genomics data. Moreover, this approach allows the identification of class-specific alterations and limits the chromosomal regions that are necessary to distinguish patients according to cancer subtype.

5.1 Introduction

In cancer, many gains and losses of chromosomes and chromosomal segments have been described. These aberrations can be detected at high resolution using the array CGH technology, widely used in oncologic research [91] and described

in Section 1.1.3. As copy number variations (CNVs) are measured within the entire genome of a disease sample compared to a normal sample [250], array CGH is ideally suited for a genome-wide identification and localization of genetic alterations involved in human diseases. An overview of algorithms for array CGH data analysis is given in [195]. Segmentation approaches divide each single-sample signal into regions of constant copy number, called segments. Subsequent classification labels the segments as gain or loss. With this consecutive approach, however, results become unsatisfactory in case of high noise levels in the data. Therefore, segmentation and identification should be performed simultaneously as these two tasks can improve each other's performance. A popular method combining both tasks is the HMM with states defined as loss, neutral, one-gain and multiple-gain. Recently, this traditional procedure was extended to a recurrent HMM in which a class of samples instead of an individual sample is modeled by sharing information on CNVs [288]. This allows identification of recurrent CNVs found at the same location in multiple samples.

In this chapter, the main focus is on ovarian cancer, which is the fourth most common cause of cancer death and ranks as the most frequent cause of death from gynecological malignancies among women in western countries [132]. In a total of 5-10% of epithelial ovarian carcinomas, a family history of breast and ovarian cancer is noted with germline mutations in mainly the tumor suppressor genes *BRCA1* and *BRCA2*. A mutation of the *BRCA1* gene cumulates the risk for ovarian carcinoma with 26-85% while a *BRCA2* mutation increases the cumulative risk with 10% [43]. There is growing evidence that *BRCA* mutated ovarian cancers have a different oncogenetic origin with distinct pathways compared to sporadic ovarian cancers. This implicates that identification of downstream genes in these pathways can lead to both a distinct clinical management for sporadic and familial ovarian cancer and new and specific targets for prevention and treatment. We therefore previously investigated genome-wide recurrent CNVs in *BRCA1* mutated versus sporadic ovarian tumors using the array CGH technology [204]. We found that losses were more frequent in the *BRCA1* tumors, probably due to the role of *BRCA1* in homologous repair of double-stranded DNA breaks. The recurrent losses were also longer in the *BRCA1* tumors (that is, 5.2Mb vs. 0.2Mb), suggesting that a huge amount of genetic material is absent in the hereditary tumors, contributing to their genetic instability. Different oncogenetic pathways were additionally demonstrated to be active in sporadic and *BRCA1*-related ovarian cancers, although larger study groups are necessary for a better comprehension of the different clinical behavior in both groups.

As array CGH data are a useful source of information, it is desirable for this type of data to be simultaneously analyzable with other high-throughput data sets. To make our kernel-based integration framework from Chapter 3 applicable to genomics data, the approach for the identification of recurrent copy number alterations is described and extended in Section 5.3. Our strategy goes beyond the

exploratory phase by using the alterations as features in a supervised classification setting and by validating these features biologically [80] [81] [82]. After applying this method to our data on ovarian cancer, we subsequently applied it to two publicly available data sets of lung cancer and oral squamous cell carcinoma to elucidate distinct pathways and to increase knowledge on tumorigenesis of these cancers (Section 5.2). This chapter ends with a description of the results in Section 5.4, followed by concluding remarks in Section 5.5.

5.2 Data

Our data were collected from patients treated for ovarian cancer at the University Hospital in Leuven, Belgium. Eight patients had a sporadic tumor without a positive family history for breast and/or ovarian cancer, while five patients were carrier of a mutation in the tumor suppressor gene *BRCA1*, involved in DNA damage repair and transcriptional regulation [313]. To make the sporadic ovarian cancer patients as similar as possible to the *BRCA1* mutated patients, only those with similar clinical and pathological characteristics were studied (that is, stage III-IV, serous papillary, and poorly differentiated). All tumor samples were collected at the time of primary surgery, with a tumor load of at least 70% as confirmed by a pathologist. DNA was extracted from the fresh frozen samples. Each tumor was hybridized twice (dye-swap) against a common reference pool. The reference cell population was extracted from peripheral blood samples of each matched participating patient. Array CGH was performed using a 1Mb BAC array CGH platform, version CGH-SANGER 3K 7, containing 3,593 clones and developed by the Flanders Institute for Biotechnology (VIB), Department of Microarray Facility and the Center of Human Genetics, Leuven, Belgium. As preprocessing, the Cy5 and Cy3 fluorescence intensities were background corrected. The ratios of Cy5 to Cy3 were subsequently calculated and normalized by the median Cy5-Cy3 ratio. Finally, these ratios were log transformed and the average of the dye-swap experiments was used as input for subsequent modeling.

In the study of Snijders *et al.* [306], a genome-wide analysis of copy number aberrations was carried out in 89 patients with oral squamous cell carcinoma, using the HumArray2.0 genome scanning array with 2,464 clones from the University of California San Francisco. Missing copy number values were imputed unsupervisedly using the s-nearest neighbors method with s set to 15 [327]. The final data set contained 2,056 unique clones. Patients were subdivided according to *TP53* mutation with 59 of them being wildtype and 16 having a mutation in the *TP53* gene. The mutation status for the remaining 14 patients was unknown. A somatic mutation causing a loss-of-function for both alleles of the *TP53* gene is one of the most common genetic alterations seen in a wide range of sporadic cancers [238].

Garnis and colleagues [134], finally, measured genome copy number profiles for 28 commonly used non-small cell lung carcinoma (NSCLC) cell lines across the whole genome using the submegabase resolution tiling array, consisting of 32,433 BAC clones. 29,781 unique clones remained after preprocessing. Twenty-two of the cell lines belonged to one of the 2 main subgroups of NSCLC: adenocarcinoma (13) and squamous cell carcinoma (9).

5.3 Methodology

As was stated in the introduction, we will use a recurrent HMM proposed by Shah and colleagues [288] for the identification of extended chromosomal regions of altered copy numbers, labeled as gain or loss. The goal of our approach is to construct features that distinguish two groups of patients and to subsequently use them in a classifier.

5.3.1 Recurrent HMM

Due to sensitivity of traditional HMMs to outliers being measurement noise, mislabeling and copy number polymorphisms within the normal human population, a robust HMM was first proposed by Shah *et al.* [289]. This version handles outliers and integrates prior knowledge about copy number polymorphisms in the analysis. To further reduce the influence of various sources of noise on the detection of recurrent copy number alterations, the robust HMM was extended to a multiple sample version in which array CGH experiments from a cohort of individuals are used to borrow statistical strength across samples instead of modeling each sample individually [288]. This makes CNVs in a small number of adjacent clones reliable when shared among many samples.

In this study, a recurrent HMM was constructed on a chromosomal basis for each group of patients separately, resulting in regions with the probability of genetic alteration among all patients per group. A clone was labeled as gain or loss when its probability of occurrence exceeded 70%. An important parameter of the recurrent HMM that needed to be set is the recurrent variable ϵ. This variable represents the probability with which samples are reflected in the recurrent CNV profile - accounting for sample-specific random effects - and is inversely proportional to the sparseness of the profile. As a trade-off between sparseness and false positive rate, ϵ was set to 0.8, although ϵ=0.7 is already sufficient for small data sets with less heterogeneity within each group of samples (that is, our data set with 8 and 5 samples, respectively). A differential region was defined as a chromosomal region that is gained/lost in one group while not being gained/lost in the other group.

Those differential regions resulting from the recurrent HMM were used as features for classification.

5.3.2 Model Building

For the data set of Snijders, a 10-fold CV strategy was applied, while LOO was chosen for the other, rather small data sets. The four different steps that had to be accomplished in each CV iteration are shown in Figure 5.1. After leaving out q samples (that is, $N/10$ for 10-fold CV; 1 for LOO), a recurrent HMM is constructed in step 1 for both groups to determine the chromosomal regions with genetic alterations that characterize each group. Combining these regions results in the chromosomal regions that are differential between the remaining $N - q$ samples from both groups. Because multiple clones can be located within each differential region, clones need to be combined. This is done per sample in the second step by taking the median of the log ratios of the clones in each region. Afterwards, a standardization is performed per sample (that is, meanshifting to 0 and autoscaling to 1) because the raw log ratios of the samples cannot be compared in absolute values. In step 3, the feature selection method DEDS (see Section 2.3) determines which preprocessed log ratios, called features, best discriminate the $N - q$ samples. The iteratively increased subset of features forms the input for classification in the last step. For all possible combinations of γ and number of features, a wLS-SVM model is built on the training set and validated on the q left out samples. This is repeated until each sample is left out once, resulting in the parameter combination with the highest AUC.

For the wLS-SVM, the linear kernel function was chosen because features in this study correspond to differential regions of consecutive genes with the same status (gain or loss), obtained with an HMM that takes the properties of CGH data into account. A kernel function that accounts for the correlation between neighboring clones and that captures the underlying categorical information in array CGH data after segmentation and identification [208] was therefore not needed and did not lead to a better performance. For the calculations in this chapter, supercomputing was needed. 5.09 years and 102,000 jobs resulted in a strategy to model array CGH data for use within a classification setting.

5.3.3 Functional Annotation Analysis

To validate the selected chromosomal regions, gene set enrichment was performed as an indication for agreement with "known" biology. Curated gene sets as defined in the Molecular Signatures Database (MSigDB) v2.1 (that is, sets of co-regulated genes from online pathway databases such as KEGG and BioCarta, publications in PubMed and knowledge from domain experts) were considered [316]. As the

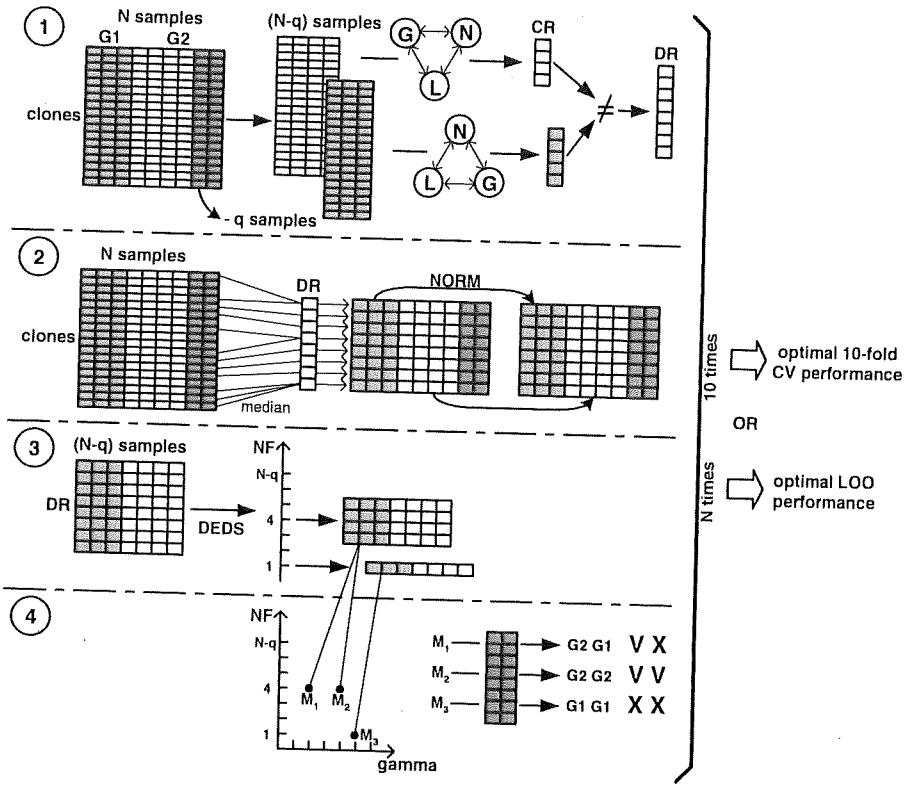


Figure 5.1: *The methodology, applied in a 10-fold CV or LOO setting, consists of four steps:* **step 1** - recurrent HMMs applied to both groups of samples, resulting in differential regions; **step 2** - conversion of clones to differential regions, and normalization (performed for each sample separately); **step 3** - feature selection using DEDS; **step 4** - wLS-SVM training on a two-dimensional grid for regularization parameter γ and the number of features, and validation on q left out samples (G1 = group 1; G2 = group 2; CR = Chromosomal Region; DR = Differential Region; NORM = Normalization; DEDS = Differential Expression via Distance Synthesis; NF = Number of Features)

HUGO gene nomenclature [354] was used in the MSigDB to uniquely characterize genes, we opted for the HUGO gene names as common identifier and used the Ensembl database [164] to extract the HUGO identifiers for the genes within the differential chromosomal regions. Those genes were subsequently divided into nine gene signatures, depending on the group (G1 versus G2 versus both) and CNV type (gain versus loss versus both). For each signature, the number of overlapping genes was calculated between all curated gene sets and the signature. For the curated gene sets, the overlap was subsequently calculated with 5000 equally-sized signatures containing genes randomly selected from the genome. The corrected method of North *et al.* [237] was used to calculate the empirical p-value for each gene set as $(r+1)/(n+1)$, with n being the number of random signatures (that is, 5000) and r the number of them with an equal or higher overlap with the gene set than obtained with the actual signature. Only gene sets with r smaller than 10 (p-value < 0.002) were investigated.

5.4 Results

Eight sporadic and five *BRCA1* mutated ovarian cancer patients were included in this study and profiled using array CGH technology. Figure 5.2 gives an impression of array CGH data that allow identification of chromosomal regions which are different in two classes of samples. It shows a recurrent amplification in *BRCA1* patients that is not present in sporadic patients. When applying the proposed methodology to this data set for distinguishing sporadic from *BRCA1* mutated patients, CNVs in 11 chromosomal regions sufficed to correctly classify 12 out of 13 samples (see Table 5.1). Table 5.2 and Figure 5.3 show information on those 11 regions, with the start and stop site determined on the full set of 13 samples. Five regions on chromosome arms 3p, 4p, 6p, 12q and Xp were gained and 3 regions on 10p, 13q and 19p were lost in *BRCA1* mutated samples. The sporadic ovarian cancer patients were characterized by loss of 3 regions on 4p and 16q. A comparison of the 11 regions found in each of the 13 LOO iterations showed a limited variability in the selected regions (Table 5.2). The 5 highest ranked features with the lowest p-value according to DEDS appeared in 8 to 11 of the 13 LOO iterations while the lower ranked regions were selected in 4 to 7 iterations.

We subsequently applied our method to two publicly available array CGH data sets for the prediction of cancer subtype. The data set of Snijders was used to distinguish patients being wildtype for the *TP53* gene from those having a mutation in *TP53*. Fifty-five of the 59 wildtype patients and 11 of the 16 *TP53*-mutated patients could be classified correctly based on 10 differential regions, none of them containing *TP53* located on chromosome 17 (see Table 5.1). The wildtype group was characterized by 3 gained regions on chromosome arms 2p, 11q and 18q, and 1 lost region on 12p. The mutated group could be distinguished with 1 gained

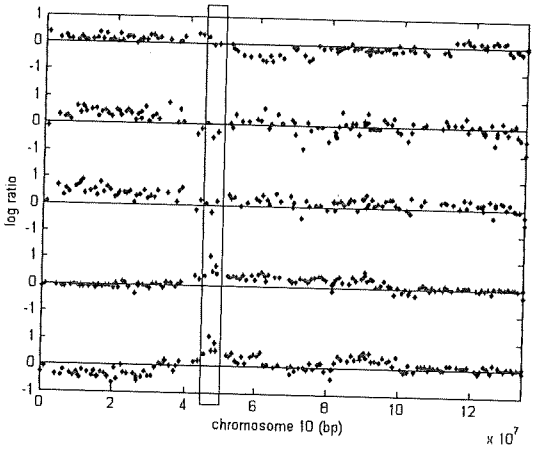


Figure 5.2: *Array CGH profile of chromosome 10 for three sporadic (top) and two BRCA1 mutated samples with ovarian cancer (bottom):* The horizontal lines indicate the 0 log ratio for all samples. Both groups have a different profile within the first $3x10^7$ base pairs and an amplification indicated by the vertical box occurs within the *BRCA1* mutated samples around $5x10^7$ base pairs.

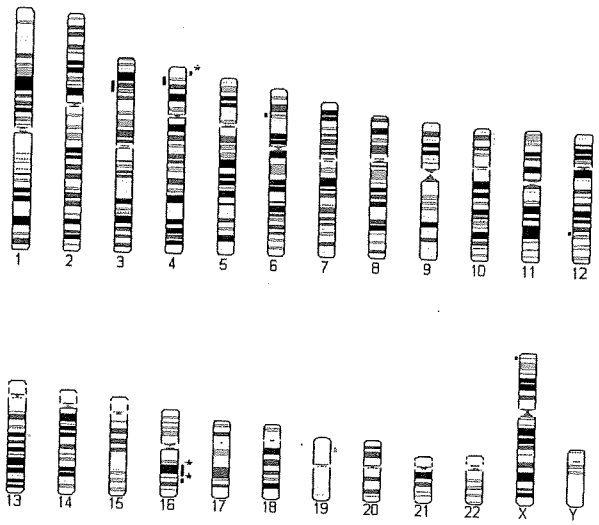


Figure 5.3: *Karyotype with the 11 differential regions selected for the prediction of ovarian cancer subtype: BRCA1* - five gained regions (shown at the left of the chromosomes) and three lost regions (shown at the right in gray); sporadic - three lost regions (shown at the right indicated with the symbol *).

Table 5.1: Performance of our modeling approach for the three considered data sets

Data set	Nb regions	Accuracy$^{\mu}$	Sensitivity	Specificity	AUC$^{\mu}$
own data	11	92.3 (12/13)	100 (5/5)	87.5 (7/8)	0.875
Snijders [306]	10	88 (66/75)	93.2 (55/59)	68.8 (11/16)	0.840
Garnis [134]	8	95.5 (21/22)	92.3 (12/13)	100 (9/9)	0.983

$^{\mu}$Represents 10-fold CV performance (Snijders) or LOO performance (own data, Garnis)

Table 5.2: Chromosomal information about the 11 differential regions selected for the prediction of ovarian cancer subtype

Feature	Chrom	Group	CNV type	Startbase	Stopbase	NG	NI
1	13q	*BRCA1*	loss	55423625	55550461	0	8
2	Xp	*BRCA1*	gain	3273880	7085387	5	11$^{\mu}$
3	12q	*BRCA1*	gain	101502349	101656438	0	9
4	4p	*BRCA1*	gain	10384154	19905375	22	11$^{\varsigma}$
5	4p	sporadic	loss	4932958	8382645	24	10
6	3p	*BRCA1*	gain	24167220	35751756	32	5
7	10p	*BRCA1*	loss	4290650	17074128	66	7$^{\varsigma}$
8	16q	sporadic	loss	56587489	67418517	81	4
9	19p	*BRCA1*	loss	12159479	13216789	39	4
10	6p	*BRCA1*	gain	24267702	29367215	86	4$^{\varsigma}$
11	16q	sporadic	loss	70089429	75199166	36	6

NG, number of genes. NI, Number of LOO iterations in which each region is selected. $^{\mu}$Region 10-50% smaller in 2 LOO runs. $^{\varsigma}$Region 10-40% smaller in 1 LOO run.

region on 10p and 5 lost regions on chromosome arms 8p, 11q and 18q. Five of these regions were selected in 8 to 10 of the 10 CV iterations while the remaining 5 regions appeared in 1 to 5 CV iterations. Loss of chromosome arms 8p, 11q and 18q in *TP53* mutated samples was mentioned in the original manuscript as well.

When applying our methodology to the data set of Garnis to distinguish adenocarcinomas from squamous cell carcinomas, 21 out of the 22 cell lines could be classified correctly based on 8 regions on chromosome arms 1q, 2p, 3q, 5p, 7q, 11p, 13q and 20p, all lost in adenocarcinoma (see Table 5.1). Five regions were selected in all LOO iterations, the regions on 7q and 11p in 19 and 15 LOO iterations, respectively, and the lowest ranked region in 9 of the 22 LOO iterations. The chromosome arms 2p and 13q were also found by Garnis and colleagues to be frequently deleted in adenocarcinoma while being amplified in the cell lines of the squamous cell carcinomas.

Because we hypothesized that genes in the selected chromosomal regions participate in processes that distinguish subtypes of cancer, a gene set enrichment-based approach was followed. This analysis highlighted enriched gene sets containing genes that may act as an oncogene or tumor suppressor gene or genes which are shown to repress tumor suppressor genes.

In our data set, the genes *BAF57* and *HOXA5* were correlated with hereditary ovarian cancer. *BAF57*, a *BRG1*-associated factor, mediates interaction with transcriptional activators or repressors, and mutation in this gene has been found to be associated with a wide variety of tumors [357]. A direct interaction is known between *BRG1*-associated factors and the *BRCA1* tumor suppressor protein. Recent studies have also shown the importance of complexes containing *BAF57* in transcriptional repression of tumor suppressor genes among which *BRCA1* [357]. Ten genes, normally down-regulated after stable re-expression of *BAF57* in breast cancer cells that lack functional BAF57, were found in four of the five regions gained in the *BRCA1* mutated samples (p-value $<$ 2e-4). In our data set, *BAF57* is thus not directly involved in the repression of the tumor suppressor activity of *BRCA1*. The gene *HOXA5* encodes a transcriptional factor with an important role in embryogenesis, hematopoiesis and tumorigenesis [55]. In human, it has been shown that *HOXA5* mRNA levels are markedly reduced or even lost in more than 60% of breast cancer cell lines and primary breast carcinoma cells. This suggests that *HOXA5* may act as a tumor suppressor gene in breast cells which makes loss of expression of this gene an important step in tumorigenesis. Six genes, normally up-regulated in cells induced by their positive regulator *HOXA5*, were found to be lost in *BRCA1* ovarian cancer (p-value = 8e-4). Loss of target genes of *HOXA5* in *BRCA1* mutated samples therefore suggests that *HOXA5*, when present, cannot accomplish its tumor suppressor role. As for the sporadic samples, loss of genes activated by *Myb* transcription factors (p-value = 6e-4) suggests a possible loss of the *v-myb* oncogene in sporadic ovarian cancer [203].

The gene *LAMA3* was important for classification of *TP53* status and was proposed by Snijders and colleagues as one of the candidate driver genes in the development of oral squamous cell carcinoma. It may play an active role in cell adhesion, migration and proliferation of the carcinoma cells [209]. Furthermore, the protein cortactin, contributing to tumor progression and metastasis formation [41], was gained in the wildtype samples while the *TP53* mutated samples were characterized by loss of the Smad transcription factors *Smad2*, *Smad4* and *Smad7*. Members of the *TGFβ/Smad* signaling pathway are potent tumor suppressor genes [267] and the inactivation of Smad factors is correlated with loss of responsiveness to *TGFβ*-mediated signal transduction [190]. Important genes lost in the adenocarcinoma cell lines were *CUTL1*, which activates a transcriptional program regulating genes involved in cell motility, invasion and extracellular matrix composition [228], and *FGF-10*, a key regulator of lung branching morphogenesis [243].

5.5 Conclusion

As increasing amounts of array CGH data become available, there is a need for algorithms to identify recurrent gains and losses based on statistically sound methods, in order to use them in classification. A large number of approaches for the analysis of array CGH data have already been proposed, ranging from mixture models and HMMs to wavelets and genetic algorithms [195]. However, in most cancer studies on array CGH data, only methods for an exploratory analysis were applied. Often a fixed threshold was used to define gains and losses, making these studies less robust against systematic changes in the baseline copy number measurements between samples [184]. An HMM on the contrary is a probabilistic method that can handle the uncertainty in the data in a formal way. This makes the HMM more robust against outliers, such as measurement noise and wrong recordings of clone locations. Moreover, we used a special variant of the HMM able to capture recurrent CNVs, by coupling the HMMs of individual samples. This turns weak CNVs into reliable features when shared among many samples. In this chapter, we used this property by first modeling CNVs in each group of patients separately. Subsequently, the alterations that were different between the two groups were used as features in a wLS-SVM model for classification. We thus propose a new methodology in which CNVs, resulting from array CGH are transformed into features for classification purposes. This general method, which is independent of cancer site, allows identifying a small set of class-specific aberrations that can distinguish patients, followed by biological validation. It can also result in clinically relevant models based on a limited set of features. In our opinion, this is one step further compared to many other studies that only perform an exploratory analysis.

The stability of the regions selected in each of the CV or LOO iterations strengthens our confidence that the chromosomal regions found with our methodology are robust. Regions lacking genes with an annotated HUGO symbol seem uninteresting at first sight. However, recent research findings on 1% of the genome indicated that 93% of the bases are transcribed, increasing the importance of non-protein-coding RNA [69]. For each data set the remaining regions were biologically validated using a gene set enrichment-based approach, although one can expect that biological validation using pathways may fail because not all genes of a certain pathway are needed for classification.

It was not self-evident to find appropriate data sets because in most studies, array CGH was used to characterize a specific set of patients without clinicopathological variables available for classification. We applied our method to two extra data sets for the prediction of disease-specific survival and distant recurrence, although not the focus in the original manuscripts [56] [124] (results not shown). Even though previous studies on CGH have shown evidence for an association of amplification with poor prognosis [5] [32], we were not able to properly classify patients according

to survival. Further research is required on the appropriateness of array CGH data in survival-related studies.

Our method performed well in distinguishing adenocarcinoma from squamous cell carcinoma. This can partly be caused by the use of cell lines instead of clinical specimens, although NSCLC cases with comparable CNV profiles have been observed. This good performance may also be ascribed to the high resolution tiling array, emphasizing the need for such arrays to strengthen our hypotheses and to refine the found regions of genetic alterations. This would increase knowledge on pathways involved in the studied cancer types. High resolution arrays thus yields potential for applicability of CNV profiles in clinical decision support.

Another approach for the classification of array CGH data has in the meantime been published [263]. The formulation of the SVM (2.6) was extended with the prior knowledge that probes should be grouped in regions and that a limited number of regions on the genome should contribute to a classifier. These assumptions were translated into constraints to enforce sparsity and similarity between successive features on the same chromosome. This prior knowledge reduces the search space and simplifies the classifier's interpretability. When compared with our approach, similar results were obtained; however, with a higher time complexity.

Chapter 6

Improved Decision Support with Graph Encoded Interactome Data

In the past, microarray studies have been criticized due to noise. Prior biological knowledge should therefore be incorporated in models based on gene expression data to improve the accuracy of diagnosis and prognosis in cancer. As prior knowledge, we investigated interaction and pathway information from the human interactome on different aspects of biological systems. By exploiting the properties of kernel methods, relations between genes with similar functions but active in alternative pathways could be incorporated in a wLS-SVM classifier. Using 10 microarray data sets, we first reduced the number of data sources relevant for multiple cancer types and outcomes. Three sources on metabolic pathway information (KEGG), protein-protein interactions (OPHID) and miRNA-gene targeting (microRNA.org) outperformed the other sources with regard to the considered class of models. Both fixed and adaptive approaches were subsequently considered to combine the three corresponding classifiers. Averaging the predictions of these classifiers performed best and was significantly better than the model based on microarray data only. These results were confirmed on 6 validation microarray sets, with a significantly improved performance in 4 of them. Integrating interactome data thus improves classification of cancer outcome for the investigated microarray technologies and cancer types. Moreover, this strategy can be incorporated in any kernel method or non-linear version of a non-kernel method.

6.1 Introduction

Although we illustrated the strength of data integration in Chapter 3, the focus in this chapter is on improving the predictive power in diagnosis and prognosis of cancer based on microarray data. In most studies, only the latter data type is available due to the maturity of the microarray technology. Microarray studies, however, have been subject to criticisms and concerns, depicted in Chapter 1, so that it would be useful to incorporate side information about gene-related interactome networks in kernel methods for classification. The earliest microarray studies were directed towards breast cancer. We refer to [239] for an overview of breast cancer signatures for which clinical phase III trials are undertaken. In these studies, however, genes were treated as single entities without regard to their neighbors in the interactome network consisting of a wide variety of interaction pairs such as protein-protein (PPI), domain-domain (DDI) and miRNA-mRNA interactions. In Figure 6.1, a simplified visualization of the mTOR pathway and its regulating pathways is depicted, representing a typical network related to oncogenesis. Several components of these pathways are deregulated in a broad spectrum of human cancers. In this example, the mTOR pathway is switched on either when expression of *RAS* is reduced due to a mutation or when the tumor suppressor gene *PTEN* is inactive. Ideally, no distinction is made between both deregulations when classifying samples as cancerous versus non-cancerous. Moreover, the importance of networks instead of individual genes has been demonstrated by the low gene overlap between prognostic gene signatures in breast cancer but high overlap in relevant pathways [376], with proliferation being the most important common driving force [97]. Genes with similar functions but active in alternative pathways should be taken into account to improve classification performance. Such complementary pathways in which a signal can be propagated through two or more parallel paths have extensively been shown to exist. A well known example shared between the majority of cancer types is *p53*-mediated apoptosis [157]. A multitude of mechanisms for apoptosis are triggered by the tumor suppressor gene *p53*, among them two distinct apoptotic signaling pathways. Independent of the specific pathway that causes inactivation of *p53*, cancer patients in which apoptosis takes place should be marked as similar with regard to the expression of those genes involved in *p53*-mediated apoptosis.

Microarray data have already been combined with biological networks; however, almost always within the gene space for the purpose of functional annotation. Vert and Kanehisa [350] extracted active pathways from gene expression data by using the metabolic network from the KEGG database as prior knowledge. Both the metabolic network and microarray data were represented as a gene x gene kernel matrix. They were combined in the feature space with a generalized form of canonical correlation analysis to learn a semantic representation independent of the two views. Hanisch and colleagues [153] proposed a variant of hierarchical clustering with an increased stability and biological plausibility of the obtained

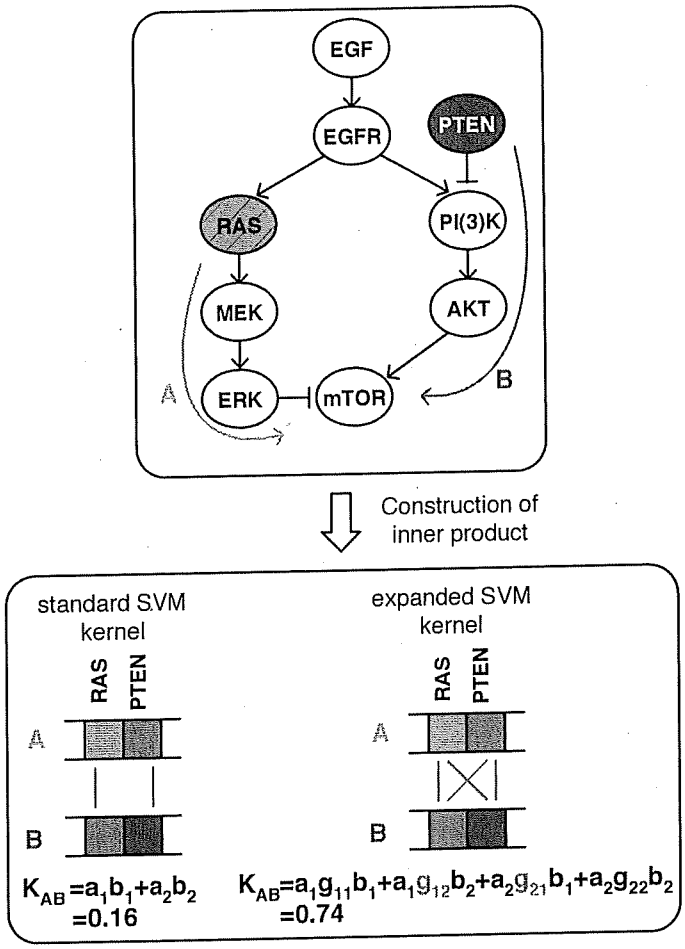


Figure 6.1: *Complementary pathway information and its incorporation in the calculation of patients' similarity:* Patients with the same phenotype can be genotypically different when alternative trajectories in a pathway are activated or repressed. Ideally, no distinction is made between alternative pathways when classifying samples. Suppose patients A and B both have breast cancer. However, patient A is characterized by repression of *RAS* while *PTEN* is silenced in patient B [230]. When building a model to distinguish breast cancer patients from control samples, the calculated similarity between patients A and B should be high. When calculating the pairwise gene-product (that is, inner product) between these patients however, they appear to be rather dissimilar with an empirical kernel value of 0.16 due to a different expression for genes *RAS* and *PTEN*. A more accurate similarity measure would be obtained by considering second order interactions between genes. Genes that are known to interact are assigned a larger weight g in the calculation of patients' similarity. Such interactions are identified based on, for example, shared pathway membership obtained from the KEGG database. This is followed by an exhaustive product between patients' expression profiles that takes gene links into account, weighted according to gene neighborhood in graphs constructed from databases such as KEGG. This will empirically increase the kernel value to 0.74.

clusters, interpretable as co-regulated pathways. This was obtained by combining a distance function derived from gene expression data and one based on a biological network. Li and Li [205] went one step further by incorporating *a priori* network information into regression analyses for the identification of genes and subnetworks related to diseases or other biological processes. They considered the Laplacian matrix of a network as penalty term in a general regression framework.

These methods all focus on genes, while the aim of our work is to improve classification for cancer patients by considering second order interactions between genes from complementary pathways. Many databases on different aspects of the interactome, of which an exhaustive overview is given in [133], are made freely available to the research community. We will refer to these databases as *secondary data sources*. Because it is useful to expand classification models with these sources, PPI networks were recently introduced to extract subnetworks of interacting proteins and to identify deregulated molecular interactions, evolving from a pathway-based to a protein-network-based approach [59] [202] [222] [321]. Rapaport and colleagues [264] applied similar ideas to differentiate irradiated from non-irradiated yeast strains, by including *a priori* pathway knowledge in the analysis of gene expression data. The high-frequency components in the data were removed with respect to the topology of the gene network, and the smoothed data, assumed to contain most biologically relevant information, were used for classification. We will not restrict our analysis to the hypothesis of similar gene expression levels for neighboring genes in a network.

The previously described approaches only considered one protein-related source. A single source, however, is not necessarily optimal for all cancer-related prediction challenges. In this chapter, we therefore do not limit our strategy to pathways or PPIs, but we will use multiple secondary data sources from databases, such as KEGG [178], REACTOME [224] and OPHID [39] (for an overview see Table 6.2). These sources contain gene-related information at other levels of biological regulation than measured with microarray technology. In order to improve microarray-based cancer classification, we investigated how to combine these sources with a patient-based kernel matrix and present a method that is able to incorporate any type of interactome data in SVM models. Where with kernel methods similarity between patients is traditionally calculated with a similarity measure based on the patients' gene expression profiles, interactome data from secondary data sources can be used to improve the way how patient similarities are calculated. At the same time, we hypothesize to improve classification performance. This is motivated by the principle that in patients within the same cancer-subgroup, different genes from the same pathway can be expressed, making single-gene markers not ideal (see Figure 6.1). Furthermore, because the relevance of the databases for each specific problem is not known beforehand and irrelevant databases may worsen the results, we present multiple well-considered schemes for combining the information from secondary data sources.

6.2 Data

Before describing how a patient-based kernel matrix can be improved with interactome data, we provide an overview of the studied microarray data sets and secondary data sources.

6.2.1 Microarray Data Sets

For this chapter, we mainly focused on breast cancer because it is one of the most extensively studied cancer types for which many microarray data sets are publicly available. In addition, data sets on ovarian cancer, prostate cancer and diffuse large-B-cell lymphoma were included. Table 6.1 gives an overview of the 16 studied data sets including at least 50 samples, with information on outcome, microarray platform and number of included samples and genes. These studies cover a wide range of predictable, cancer-related outcomes such as response, relapse, metastasis and survival. For the data set of Bild [29], no binary outcome was available. The median survival time of 3 years was chosen as cut-off to balance both classes of samples. For the data set of Rosenwald [273], the suggested 4 years was taken while in the study by Wang [359], metastasis within 5 years was studied. For these data sets, censored samples with last follow-up before the chosen threshold (3, 4 or 5 years, respectively) were excluded, resulting in the loss of 14, 20 and 10 samples, respectively.

The data sets gathered with the Affymetrix microarray technology were preprocessed with MAS5.0. An updated array annotation was used for the conversion of probes to Entrez Gene IDs [89]. In [273], a custom-designed microarray composed of genes of which the products are preferentially expressed in lymphoid cells was used. For this case, the data at the Entrez Gene level as provided by the authors were used and missing gene expression values were imputed unsupervised using the s-nearest neighbors method with s set to 15 [327]. We also took the low signal-to-noise ratio of microarray data into account by unsupervised exclusion of genes with low variation. The 5,000 most varying genes were retained. Limiting the number of genes to 5,000 has the additional advantage that the computation of the pseudoinverse Laplacian matrix remains tractable (see Section 6.3.1), although the analysis should be performed on a sufficient number of genes between which interactions are described in secondary data sources.

Table 6.1: Microarray data sets

	Data set	Cancer type	Outcome	Platform	nb samples (neg/pos)	nb genes
T1	Berchuck [23]	OC	Binary survival (short $\leq$ 3 yrs vs. long $>$ 7 yrs)	U133a	53 (29/24)	12,633
V1	Bild [29]	OC	Binary survival (short $\leq$ 3 yrs vs. long $>$ 3 yrs)	U133a	133 (88/45)	11,911
V2	Chin [56]	BC	Distant recurrence (no vs. yes)	U133av2	129 (102/27)	12,633
T2	Hess [159]	BC	Pathologic response (RD vs. CR)	U133a	133 (99/34)	12,633
V3	Huang 1 [163]	BC	Disease recurrence (no vs. yes)	U95av2	52 (34/18)	8,740
V4	Huang 2 [163]	BC	Relapse (no vs. yes)	U95av2	80 (53/27)	8,740
T3	Ivshina [171]	BC	Local, regional or distant recurrence (no vs. yes)	U133a+b	249 (160/89)	18,001
V5	Miller [229]	BC	Death from BC (no vs. yes)	U133a+b	236 (181/55)	18,001
T4	Pittman 1 [251]	BC	Relapse (no vs. yes)	U95av2	158 (95/63)	8,740
V6	Pittman 2 [251]	BC	Loco-regional recurrence (no vs. yes)	U95av2	158 (132/26)	8,740
T5	Pittman 3 [251]	BC	Distant metastasis (no vs. yes)	U95av2	158 (108/50)	8,740
T6	Rosenwald [273]	DLBCL	Overall survival (short $<$ 4 yrs vs. long $\geq$ 4 yrs)	Lymphochip	220 (118/102)	6,707
T7	Singh [304]	PC	Tumor status (normal vs. tumor)	U95av2	102 (50/52)	8,193
T8	Sotiriou 1 [310]	BC	Relapse (no vs. yes)	U133a	187 (120/67)	12,633
T9	Sotiriou 2 [310]	BC	Distant metastasis (no vs. yes)	U133a	179 (139/40)	12,633
T10	Wang [359]	BC	Metastasis within 5 yrs (no vs. yes)	U133a	276 (183/93)	11,911

T, training data set; V, validation data set. BC, breast cancer; OC, ovarian cancer; DLBCL, diffuse large-B-cell lymphoma; PC, prostate cancer. RD, residual disease; CR, complete response

6.2.2 Secondary Data Sources

The human interactome is the compendium of all stable, transient, direct and indirect physical interactions between proteins in molecular machines and pathways in an active cell. For cancer, much biological knowledge and pathway information is available in databases on different aspects of biological systems [49] [133]. A distinction can be made between protein-protein interactions (PPI), domain-domain interactions (DDI), metabolic pathways, signaling pathways, transcription factors (TF), gene regulatory networks and protein-compound interactions. Table 6.2 gives an overview of the considered secondary data sources, with the type of pathway information contained in each of them, the definition of links between genes, and the number of extracted gene pairs. These data sources are used as input in our methodological setting, presented in Figure 6.2.

Metabolic Pathways

Metabolic pathways, which are series of biochemical reactions with focus on chemical modifications, were extracted from four databases as each database may contain false positive pathway predictions. Moreover, because most metabolic reactions are linked with disease genes, the human metabolic network is rather fragile [215]. KEGG [178] stands for Kyoto Encyclopedia of Genes and Genomes and is a knowledge base for the analysis of gene functions in terms of networks of genes and molecules. We focused on the KEGG biochemical pathway database containing metabolic pathways, pathways related to genetic information processing, regulatory pathways involved in environmental information processing and cellular processes, and pathways active in human diseases and drug development. As the pathway information is matched with the KEGG gene database, individual genes can be linked to components of KEGG biochemical pathways. Based on these cross-references, we extracted gene pairs defined as genes encoding for proteins that can catalyze two reactions in the same pathway.

Another large repository on metabolic pathways is BioCyc, a collection of more than 350 organism-specific pathway/genome databases (PGDBs) for most eukaryotic and prokaryotic species with sequenced genomes [50]. Every BioCyc PGDB contains the predicted metabolic network for one organism, including metabolic pathways, enzymes, metabolites and reactions. These databases can be divided into three groups, according to the amount of manual curation. In this chapter we considered two PGDBs, the intensively curated database MetaCyc [50] and the computationally-derived database HumanCyc, subject to a moderate curation of less than 1 year [272]. HumanCyc contains information on 28,783 human genes, their products and the metabolic reactions and pathways they catalyze. The metabolic pathways were predicted based on genome

annotation, with missing enzymes within the predicted pathways replaced by candidate proteins when possible. MetaCyc [50], the Multiorganism database of Metabolic Pathways and Enzymes provides a high-quality resource on small-molecule metabolisms and contains experimentally verified metabolic pathway and enzyme information curated from specialist literature. We extracted the human pathways from MetaCyc; however, only one metabolic pathway was not included in HumanCyc. Gene pairs were in analogy to KEGG defined as genes with proteins belonging to the same metabolic pathway.

Because not all known human metabolic pathways are included in HumanCyc and KEGG, we also considered a more recent human metabolic network, manually reconstructed by integrating genome annotation information from different databases and metabolic reaction information from specialist literature [215]. This network, referred to as EHMN (Edinburgh human metabolic network) contains more than 2,000 metabolic genes and almost 3,000 metabolic reactions, reorganized into 70 human-specific metabolic pathways. Also here, gene pairs were defined as genes encoding for enzymes that are involved in the same reaction.

The fourth secondary data source is REACTOME [224], an expert-authored knowledge base of human biological processes. The database consists of 2,907 reactions involving 2,975 human proteins and grouped into pathways taking their temporal relationships and interdependencies into account. It represents pathways of intermediary metabolism, regulatory pathways, signal transduction and high-level processes such as the cell cycle. Gene pairs were defined as genes involved in the same reaction or complex.

Protein-Protein Interactions

Only a limited number of proteins are able to physically attach to a huge number of proteins [16]. PPI networks are therefore scale-free, with the number of nodes following a power law distribution in function of the number of interactions. Several major databases on PPIs are available. PPIs are typically extracted from publications in which high-throughput proteomic or small-scale biomolecular methods were applied. Especially high-throughput methods make PPI networks noisy with many false positives and inaccurate with inconsistent annotations. In addition, each available PPI database uses its own extraction, curation and storage protocols, they do not necessarily explore the same scientific papers and are composed of different compositions of experimentally and computationally determined interactions. Prieto and De Las Rivas [257] have shown a limited intersection and overlap between the six major PPI databases for human proteins (BioGRID, BIND, MINT, HPRD, IntAct, DIP) [12] [52] [10] [180] [279] [38]. As the information contained in these databases is partly complementary, knowledge about the interactome can be increased and improved

by combining multiple databases. Notably, even the union of all databases is still incomplete with many unknown components and pathways, reaching coverage of 31% of the human proteome, corresponding to 12,053 proteins and 83,670 interactions. The five largest databases are considered in this chapter, with HPRD containing 63.3% of the known PPIs, BioGRID 40.9%, IntAct 34.9%, MINT 22.5% and BIND 9.7% according to data in 2008 (see the 'statistics' section on http://bioinfow.dep.usal.es/apid/, [257]). We did not consider DIP as this database only contains 1.7% of the known PPIs [279].

Four of the largest databases (BIND, HPRD, MINT and IntAct) are combined with high-throughput experiments in the Online Predicted Human Interaction Database (OPHID), a catalog of 60,675 known human PPIs [39]. This manually-curated, literature-derived catalog was further expanded with predictions for 34,824 interactions that occur in other model organisms but with both human orthologs conserved in humans. In this way, OPHID could extend the human interactome with a set of proteins that were not yet included in literature-based databases. We translated all biomolecules included in OPHID to one or multiple corresponding genes, after which relations between the resulting genes were extracted. The same definition for gene pairs was used for the remaining PPI database BioGRID [38]. As this database is multi-organismal, we only extracted the interactions between human genes. The list of gene pairs for this database was much smaller than with OPHID because the conversion from protein to gene had already been done by its curators.

STRING [174], the Search Tool for Recurring Instances of Neighboring Genes, is extracted from PPI networks and predictions based on comparative genomics and text-mining. It lists potential functionally associated proteins based on the genomic association of their genes. It thereby acts as a metadatabase that maps all interaction evidence onto a common set of genomes and proteins, by weighting and integrating information from numerous sources: PPI databases, high-throughput experimental interaction data, associations highlighted in published literature, interaction transfer between organisms, functional co-expression of genes, conserved neighborhood, gene fusions and phylogenetic co-occurrences. All resulting individual scores are integrated in a combined score, reflecting the confidence for each predicted protein association. This results in an increased confidence when an association is supported by multiple types of evidence. We truncated the interaction network to a stringency of 500 for the confidence score and only retained gene interactions that were experimentally determined.

Domain-Domain Interactions

As often only a fraction of a protein directly interacts with its biological partners, we also investigated the use of inter-chain DDIs, that is, interactions between

domains from different proteins. To increase coverage and quality, we considered two comprehensive resources that have collated all known and predicted DDIs from various sources. The database DOMINE [261] and the Unified Domain Interaction database (UniDomInt) [31], both using Pfam A domain definitions, combine DDIs based on experimentally derived three-dimensional structures in the Protein Data Bank and DDIs predicted by 8 and 9 computational approaches, respectively. In the DOMINE database [261], each predicted DDI is given a discrete confidence level high, medium or low. We did not include the low confident DDIs that were only predicted by one computational method, reducing the number of DDIs to 6,689. UniDomInt [31], on the other hand, provides a continuous reliability score between 0 and 1. We set the minimal threshold to 0.1, including 8,470 DDIs. Genes of which the proteins interact via a domain-domain interaction were linked in the corresponding graph representations.

Protein Families and Domains

To cover an as wide as possible range of gene-related information, we additionally investigated whether incorporating information on proteins belonging to the same protein family or sharing the same protein domain improves classification. Information on the structural and functional properties of proteins was extracted from PROSITE [299] and Pfam [119]. PROSITE is a database that uses amino acid patterns and profiles for the identification of protein families, domains and functional sites. The raw data provided by KEGG were used to extract gene pairs with one or multiple protein domains, families or functional sites in common. Pfam [119] is another large comprehensive and accurate collection of protein domains and families. Using the same method for extraction as for PROSITE, no distinction could be made between the manually curated Pfam A families and the automatically generated Pfam B families. Also for Pfam, gene pairs were defined as genes with proteins belonging to the same protein family or sharing the same protein domain.

Transcription Factors

Finally, evidence has suggested that miRNAs may play a role in human cancers [212]. We therefore considered four available miRNA databases. The miRBase (microRNA database) [147] contains a pipeline for predicting miRNA target genes in mRNA sequences based on the miRanda algorithm. P-values are assigned to individual miRNA-target binding sites, connecting each miRNA to a list of predicted gene targets. For the 851 miRNAs included in the database, we set the p-value to 0.001 to include only the most confident predicted miRNA-target assignments. MiRNAmap [162] is another database containing 470 miRNAs, thereby distinguishing known miRNA genes obtained from miRBase and putative

miRNA genes identified by comparative sequence analysis. Besides experimentally verified miRNA targets obtained from both specialist literature and TarBase [287], three computational tools miRanda, RNAhybrid and TargetScan are used for the identification of putative miRNA targets. To reduce the rate of false positive target site predictions, we applied the three criteria that were proposed by Hsu and colleagues [162]: target sites must be predicted by at least two tools, they must be located in accessible regions, and the target genes must contain multiple target sites. Next, microRNA.org [24] contains target predictions also based on the miRanda algorithm with miRNA sequences obtained from miRBase, but the predicted target sites of 677 miRNAs are filtered according to evolutionary conservation of sequence blocks across multiple vertebrates. We set the threshold for the conservation score to 0.70 to select target sites that are conserved in mammals. Finally, the target site prediction tool TargetScan [126] provides a p-score, corresponding to a Bayesian estimate of the probability that a site is conserved due to miRNA targeting. These scores reflect the biological relevance and efficacy of each site. We set the threshold to 0.3, corresponding to 74 miRNA genes. For all these databases, gene pairs were defined as genes targeted by the same miRNA.

The programming language Perl was used for the conversion of each secondary data source into a list of gene pairs, used as input for the calculation of the pseudoinverse Laplacian.

Table 6.2: Secondary data sources

Secondary data source	Type of pathway resource	Definition of a gene pair	Nb gene pairs	Release	Reference
1. KEGG	metabolic pathways	genes of which the proteins belong to the same pathway	609,269	49.0	[178]
2. HumanCyc	metabolic pathways	genes of which the proteins belong to the same pathway	12,314	12.0	[50] [272]
3. EHMN	metabolic pathways	genes of which the proteins belong to the same pathway	198,876	/	[215]
4. REACTOME	metabolic pathways, signaling pathways	genes involved in the same reaction or complex	722,508	28	[224]
5. OPHID	protein-protein interactions, genetic interaction networks	genes of which the proteins interact	221,674	1.71	[39]
6. BioGRID	protein-protein interactions	genes of which the proteins interact	40,812	2.0.54	[38]
7. STRING	protein-protein interactions	experimentally determined gene interactions	315,686	8.1	[174]
8. DOMINE	domain-domain interactions	genes with proteins interacting via a domain-domain interaction	18,213,973	1.1	[261]
9. UniDomInt	domain-domain interactions	genes with proteins interacting via a domain-domain interaction	20,506,327	Aug '09	[31]
10. PROSITE	protein families and domains	genes with one or multiple protein domains, families or functional sites in common	6,267,453	20.0	[299]
11. Pfam	protein families and domains	genes with one or multiple protein domains or families	3,649,554	23.0	[119]

Continued on next page

Table 6.2 – continued from previous page

Secondary data source	Type of pathway resource	Definition of a gene pair	Nb gene pairs	Release	Reference
		in common			
12. miRBase	transcription factors, gene regulatory networks	genes targeted by the same miRNA	6,819,380	5	[147]
13. miRNAmap	transcription factors, gene regulatory networks	genes targeted by the same miRNA	28,120,039	2	[162]
14. microRNA.org	transcription factors, gene regulatory networks	genes targeted by the same miRNA	13,207,828	Sept '08	[24]
15. TargetScan	transcription factors, gene regulatory networks	genes targeted by the same miRNA	8,580,619	5.1	[126]

6.3 Methodology

In this section, we first introduce the Laplacian from spectral graph theory. This concept is then used to improve the way how patient similarities are calculated. Because the relevance of each secondary data source is not known beforehand, we also present multiple combination schemes.

6.3.1 Pseudoinverse Laplacian

Many biological processes are representable as a large-scale sparse network. Each secondary data source, and more specifically the list of extracted gene pairs, can therefore be represented as a weighted, undirected graph with symmetric weights $w_{kl} \geq 0$, assigned to each edge between a pair of different nodes k and l. Such a graph is composed of p nodes representing the genes, and the edges connect genes that are linked with regard to the secondary data source under study. This graph is characterized by a weighted adjacency matrix $W = [w_{kl}]$, $k, l = 1..p$, and the diagonal degree matrix D with degrees d_1 to d_p as diagonal elements. The degree of a node k is defined as the sum of the weights w_{kl} for node k across all nodes l. From spectral graph theory, we can now define the unnormalized graph Laplacian matrix L as the difference between the degree matrix and the weighted adjacency matrix ($L = D - W$). This matrix is symmetric and positive semidefinite. Genes belonging to similar pathways will be connected by a relatively large number of short paths, while fewer, typically longer paths connect genes with completely separate functions. Finally, as our aim is to add extra gene-related information with respect to the traditional similarity measure rather than to discard genes for which no information is available, self-loops were added to isolated genes, setting their degree to 1.

Fouss and colleagues [122] have shown that the Moore-Penrose pseudoinverse $L^\dagger$ of the Laplacian matrix of a graph can be interpreted in terms of similarity between pairs of genes in the interacting network. Matrix entries increase when the number of paths connecting two nodes increases and when the length of the paths decreases. In this chapter, we will present the p x p pseudoinverse Laplacian matrix $L^\dagger$ of the graph corresponding to secondary data source m as $G_m = [g_{kl}]$ $k, l = 1..p$. The graph of each secondary data source and the extraction of G_m are presented as step 1 in Figure 6.2. In case of a fully connected graph, the G-matrix is expected to be fully dense. Isolated genes, however, introduce zero patterns in the inverse, thereby reducing the density as will be illustrated in Table 6.3. Other graph kernels such as the diffusion kernel [187] are less suitable since they require the optimization of an extra parameter. Also, instead of the unnormalized Laplacian, the normalized version $D^{-1/2}LD^{-1/2}$ in which the connectivity of each gene is taken into account can be considered. However, the use of the normalized Laplacian did not lead to an improvement in performance (results not shown).

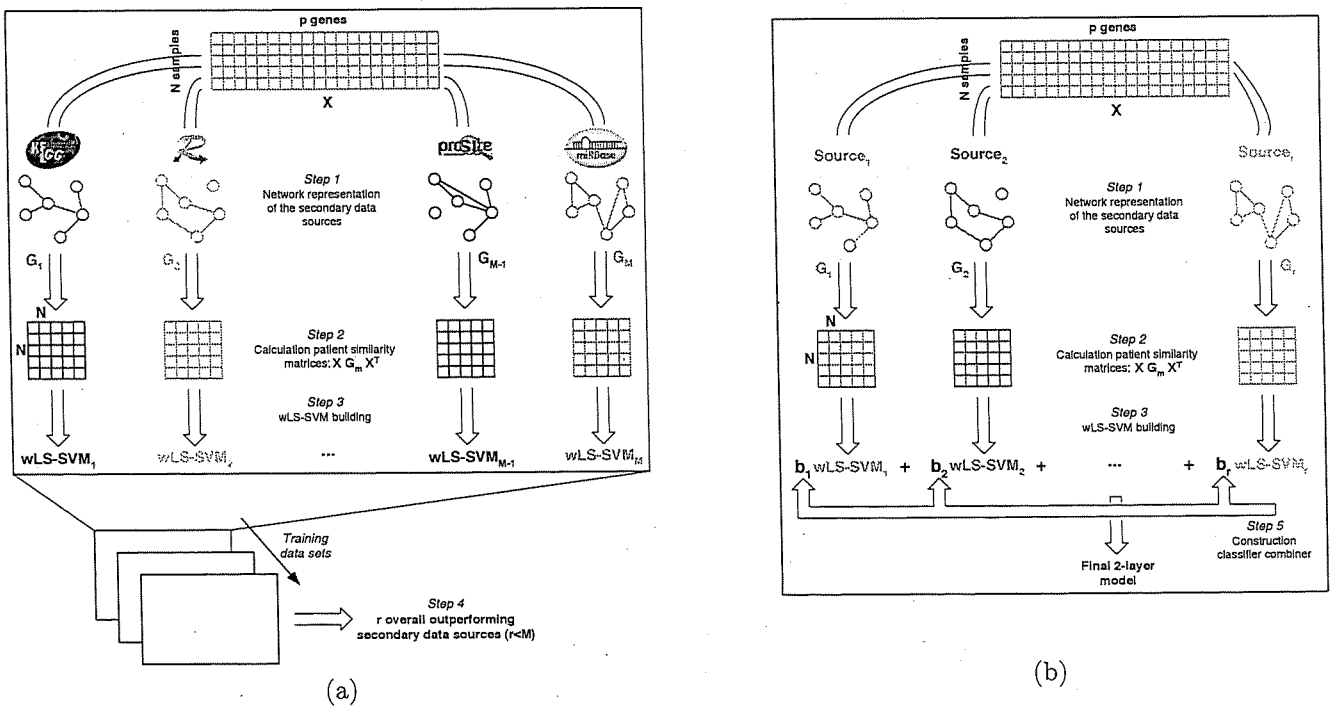


Figure 6.2: *Global overview of the methodology:* (a) In step 1, M secondary data sources are represented as a graph. This graph-related information is subsequently incorporated via the pseudoinverse of the Laplacian (the G-matrix) into the calculation of patient similarity. In step 3, wLS-SVM models are built on each of the updated kernel matrices obtained in step 2. Based on several microarray training sets, r secondary data sources that increase performance for all training data sets with respect to models built on only microarray data are selected out of the M sources. (b) After having selected the r outperforming secondary data sources, steps 1 to 3 are repeated for those r sources. In step 5, a classifier is learned and constructed for the combination of the r corresponding wLS-SVM models.

6.3.2 Adapted Kernel Function

A kernel matrix based on a linear kernel function is represented as $K = XX^T$, with $X = [x^{1^T}; ...; x^{N^T}]$ the N x p patient microarray data, N the number of samples, and p the number of measured genes (here, reduced to the 5,000 most varying genes). In patient domain, each matrix entry K_{ij} corresponds to $x^{i^T}x^j$, with x^i and x^j the gene expression profiles of patients i and j, respectively. We incorporated a secondary G-matrix in our kernel-based classification framework by expanding the kernel matrix XX^T to XGX^T (step 2 in Figure 6.2). Each entry in this expanded kernel matrix now corresponds to $K_{ij} = \sum_{k,l=1}^{p} x_k^i g_{kl} x_l^j$. Each G-matrix thus exhaustively relates the gene expression profiles of patients, weighted by its entries g_{kl}. As each secondary data source leads to a different kernel matrix, normalization is required to make them comparable. The normalized kernel matrix $\widetilde{K}$ was therefore considered, defined as $K/trace(K)$ with $trace(K)$ the sum of the diagonal elements of K. The proposed kernel function with incorporation of secondary data can easily be extended to out-of-sample data points as the G-matrices with gene-related information do not need to be recomputed. This allows easily validating and testing the obtained models.

6.3.3 Combining Classifiers

It is not known beforehand which secondary data sources are relevant for the problem at hand and thus which of the sources will increase prediction accuracy. However, combining the output of multiple classifiers may improve overall prediction accuracy. For this combination, a distinction can be made between the following architectures [172]: a *parallel* architecture by which the results of classifiers are combined, possibly after classifier selection or weighting by a gating device; a *cascading* architecture with a linear sequence of classifiers; a *hierarchical* architecture with a decision tree structure. We opted for a parallel structure and combined at a second level the predictions of the wLS-SVM models, each built with the inclusion of one secondary data source (step 5 in Figure 6.2b). Duin and Tax [103] have shown that no combining rule is optimal in multiple combination tasks. We therefore considered multiple combination schemes that have proven to be useful in specialist literature and that provide continuous predictions on which the AUC can be calculated.

A distinction is made between fixed and trained combining rules [102] [103]. As fixed combining rule, we investigated mean and median as they are most appropriate for the combination of similar outputs. This is likely to be the case in this set-up as the latent variables of the wLS-SVM are obtained from the same microarray data set, modified by a specific secondary data source. Fixed combining rules, however, are almost always suboptimal while a trainable combiner may lead to a more significant improvement with respect to static combiners [102].

Moreover, when a linear or nonlinear combiner is trained, statistically independent classifiers are not needed [183]. Although requiring additional data and at the cost of extra training, a weighted sum of the predictions of the individual classifiers was considered. The weights were set to the raw training AUC values or to these values after scaling to the half open interval]0,1]. Furthermore, an exhaustive search among all possible unweighted combinations of individual classifiers was performed. We additionally considered three more advanced models, being naïve Bayes (NB), logistic regression (LR) and linear discriminant analysis (LDA). Both NB and LDA are ideal methods when the number of training observations is limited. Although NB is based on the assumption of independence between predictors, a good performance has been shown for functional dependencies, that is, predictors that are generated from the same underlying distribution [268].

We also investigated a combination at the level of the kernel matrices, by alternatingly optimizing the weights assigned to the kernel matrices and the parameters of the wLS-SVM classifier. With this approach however, only one kernel matrix was assigned a weight of 1 while all other matrices obtained the weight 0. This led to rather poor performances due to the selection of one kernel matrix for building the final wLS-SVM classifier. As we were interested in an improved performance rather than in the selection of one specific interactome database, we opted for the ensemble approach.

6.3.4 Model Building

In this study, each data set was randomly split into 10 folds, stratified to outcome. 50% of the data corresponding to folds 1 to 5 was used for training the individual classifiers. This part of the data was normalized per gene, and the obtained gene characteristics were used for the normalization of folds 6 to 10. An internal 5-fold CV on folds 1 to 5 was used for the optimization of the regularization parameter γ, and the wLS-SVM model with the optimal γ was rebuilt on the entire training data. Similar results were obtained when the L-curve was used for the selection of γ [156]. As the γ values of the individual wLS-SVM models were checked cautiously to prevent overfitting and to assure good generalization performance, the combining rules were learnt on 80% of the data of which 30% (that is, folds 6 to 8) were new with respect to the first training phase. Considering 80% of the data in the second phase for training a combined classifier has the extra advantage that the peaking phenomenon is countered, defined as the decrease in classification accuracy when too many features are included in the classifier [301]. Among the considered combined classifiers, especially LDA suffers from this phenomenon [265]. LDA is also poorly posed when the number of observations and parameters to be estimated are comparable [125]. The use of 80% of the data guarantees that the number of training observations sufficiently exceeds the number of individual classifiers. The last 2 folds corresponding to 20% of the data were used for validation. Not only the

combined classifiers but also the baseline wLS-SVM model built only on microarray data and the models with the individual use of each of the secondary data sources were validated on the same observations. The split of the data into 10 folds was repeated 200 times. The AUC values of the baseline model were compared with the ones of all other models using the one-sided paired-sampled t-test, with 0.05 as significance level. However, the overlap in training and test set between the 200 experiments can increase the probability of type I errors (that is, rejecting a true null hypothesis). Moreover, we were not only interested in the performance for a specific problem, but in the general performance for multiple tasks. The obtained results per data set were therefore confirmed by repeating the comparisons over all data sets with the average performance per set, using the non-parametric Wilcoxon signed ranks test [95].

To assess the dependency between individual wLS-SVM models, we considered mutual information (MI) [131]. When two random variables x and y have probability distributions $P(x)$ and $Q(y)$, MI is defined as the relative entropy between the joint distribution $R(x, y)$ and the product distribution $P(x)Q(y)$: $MI(x, y) = \sum_{x,y} R(x, y) log_2(R(x, y)/[P(x)Q(y)])$. In other words, MI represents the reduction in uncertainty about one variable given the other, meaning that variables are statistically independent when MI=0. A higher MI indicates that the two variables are non-randomly associated with each other [131]. In this study, the variables x and y were binary and defined as 1 when the classifier predicted a sample to be of the positive class and 0 otherwise.

6.4 Results

In this chapter, we present a framework for microarray-based cancer classification in which the secondary data sources described in Section 6.2.2 can be incorporated. These databases were converted into graphs from which the corresponding Laplacian matrix could be derived (step 1 in Figure 6.2). To represent similarity between pairs of genes, we used the pseudoinverse of the Laplacian [122], referred to as G-matrix. With this graph-based approach, both direct and indirect connections between genes, and thus their neighborhood in the human interactome are taken into account. Genes that do not belong to the same pathway but are connected to a same subset of genes, for example, are assigned a positive coefficient in the secondary G-matrix. These matrices were incorporated in the calculation of patient similarity (step 2 in Figure 6.2). This corresponds to replacing the standard inner product by a quadratic form defined upon G^{-1} ($x^T G^{-1} x$), and can be interpreted as a weighted version of the standard inner product. An example is given in Figure 6.1.

Table 6.3: Characteristics of the Laplacian and G-matrices

Secondary data source	Laplacian: median nb gene pairs	Laplacian: 25^{th}-75^{th} percentile	G-matrices: median nb gene pairs	G-matrices: 25^{th}-75^{th} percentile
KEGG	69,870	41,771 - 99,310	5,553,761	4,950,232 - 6,473,276
HumanCyc	533	298 - 752	4,616,183	4,548,199 - 5,524,069
EHMN	5,416	3,339 - 11,847	9,356,394	8,954,741 - 10,040,781
REACTOME	43,611	27,716 - 59,786	6,453,263	4,585,327 - 7,725,010
OPHID	5,261	2,783 - 9,601	10,821,416	9,713,765 - 11,068,831
BioGRID	1,995	1,067 - 3,613	11,047,443	9,744,897 - 11,811,681
STRING	4,327	2,531 - 8,512	10,563,906	10,331,629 - 10,811,510
DOMINE	744,622	568,318 - 878,721	11,221,950	10,719,122 - 11,353,931
UniDomInt	863,228	667,232 - 1,019,495	11,541,609	10,635,969 - 11,826,811
PROSITE	256,028	221,618 - 276,509	10,967,586	10,962,602 - 11,560,607
Pfam	145,582	112,127 - 153,525	12,199,330	12,014,816 - 12,258,684
miRBase	342,756	275,376 - 347,322	10,136,253	9,243,720 - 10,163,274
miRNAmap	430,310	237,475 - 600,733	2,554,930	1,517,502 - 3,324,331
microRNA.org	228,423	148,857 - 350,045	2,025,078	1,410,494 - 2,800,161
TargetScan	422,876	242,981 - 656,470	5,420,278	3,766,078 - 6,798,828

To get an idea about the extra information that is added by the secondary data sources with respect to the standard inner product, Table 6.3 shows the median (25th-75th percentile) number of second order interactions included in each of the 15 secondary data sources. These numbers slightly vary per source as the 5,000 most varying genes selected from each microarray data set differ. The larger amount of second order interactions in the G-matrices with respect to the Laplacian matrices shows that the network neighborhood is captured by taking the pseudoinverse of the Laplacian.

6.4.1 Selection Outstanding Secondary Data Sources

Although the idea of secondary data incorporation is applicable to any kernel method or method that can be kernelized, we present results for the wLS-SVM model. To reduce the set of considered secondary data sources to the ones that are relevant across multiple cancer types and usable for multiple outcomes (represented in Figure 6.2a as step 4), 10 out of 16 microarray data sets were randomly selected for training, indicated in Table 6.1. The 10-fold CV strategy depicted in Section 6.3.4 was repeated 200 times to obtain a good estimate of the generalization error. In order to objectively select the secondary data sources that performed well for the training microarray data sets, models built on an individual secondary data source with a better average performance than the baseline model (without inclusion of secondary data) were given a decreasing score starting from 15 for the best model. Models that performed worse than baseline were given the score 0. Adding those scores for each secondary data source over all training sets highlighted three outstanding secondary data sources: KEGG, OPHID and microRNA.org (see Table 6.4). Only these three databases were used for the remaining data sets.

The mutual information between the models based on one of the three selected secondary data sources was on average 0.376, varying from 0.174 to 0.607 for the 10 training data sets and indicating a certain degree of independence between the three sources. No relationship was found between MI and the increase in performance with respect to baseline. Seven of the 10 training microarray data sets are from breast cancer (see Table 6.1). However, as can be seen in Table 6.4, the increase in performance caused by the three selected secondary data sources was not limited to breast cancer. KEGG performed well for all cancer types, OPHID resulted in a better performance for breast cancer, ovarian cancer and lymphoma, while microRNA.org led to an increased performance for both breast and prostate cancer. Based on the training data sets, the data sources STRING and TargetScan seemed to be specific to breast cancer. Finally, there were no databases that simultaneously led to an improved performance for the same training data sets.

Table 6.4: Selection of secondary data sources based on the training microarray data sets

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
T1	14	9	0	11	10	0	0	13	15	12	0	8	0	0	0
T2	0	0	0	0	14	0	15	0	0	0	0	0	0	0	0
T3	0	0	0	14	15	0	0	0	0	0	12	0	11	13	10
T4	15	0	0	0	11	0	13	0	0	0	0	0	12	14	0
T5	15	0	0	0	0	0	11	0	0	0	0	0	14	12	13
T6	14	13	0	0	15	0	0	0	0	0	0	0	12	0	0
T7	15	0	0	14	0	0	0	0	0	0	0	0	0	13	0
T8	0	13	0	0	0	0	0	0	0	0	0	0	0	14	15
T9	7	0	12	0	15	9	10	14	0	11	0	0	0	13	8
T10	0	0	0	0	12	0	0	0	0	15	0	0	14	13	0
Sum	**80**	35	12	39	**92**	9	49	27	15	38	12	8	63	**92**	46

Individual models with a better performance than the baseline model are given a decreasing score starting from 15 for the best model; models that perform worse are given the score 0.

6.4.2 Combination Multiple Classifiers

Because the relevance of the three sources for each specific classification task is not known beforehand, the three corresponding individual classifiers were combined at a second level (step 5 in Figure 6.2b). Three types of combination rules were considered: 1) fixed rules for which no training is required (that is, mean and median), 2) simple trained rules for which the influence of each model on the final prediction is determined by their individual training performance or for which the optimal combination of individual models is obtained with an exhaustive search, and 3) more advanced models, being naïve Bayes, logistic regression and linear discriminant analysis. For each of the 200 experiments, the last two types of combination rules were trained on 8 folds (that is, including the 5 folds upon which each model was built expanded with 3 untouched folds) and validated on the remaining 2 folds. The global workflow consisting of the graph representation of the secondary data sources, the incorporation into the calculation of patient similarities, model building, the selection of the relevant secondary data sources, the combination of relevant models and validation, is provided in Figure 6.2.

6.4.3 Training and Test Results

The results for the 10 training microarray data sets are shown in Figure 6.3 and Table 6.5. The five bars per data set represent the mean AUC values for the following models: 1) the baseline model built on the microarray data only, 2) the model based on the secondary data source with the best performance (KEGG, OPHID or microRNA.org), 3) the combination of models according to the best fixed rule, 4) the combination according to the best trained rule, and 5) the combination of individual models using the best advanced approach. Figure 6.4a provides an overview of the performance of the three selected secondary data sources for all training data sets. As the prediction accuracy was evaluated by applying our 10-fold approach 200 times, the 200 test AUC values of the baseline model were compared with the 200 test AUC values of the other considered models using the one-sided paired-sampled t-test. The p-values for these comparisons are shown per data set in Table 6.5. These results have been confirmed by applying the Wilcoxon signed ranks test to the average performance of all data sets. The models based on the best individual secondary data source outperformed the baseline models over all training data sets (p-value 0.002). Also combining classifiers based on individual secondary data sources with a fixed or trained rule outperformed the baseline models with a p-value of 0.0039 and 0.0098, respectively. The more advanced models did not perform unambiguously better (p-value 0.557), although an improvement was observed when reducing the number of classifiers.

To define the most optimal combination rules, we compared the average performance for the fixed, trained and advanced rules in Figures 6.4b to 6.4d, respectively. Based on the training data sets, the mean fixed rule, the weighted trained rule and naïve Bayes performed best. Only these three rules were therefore validated on the remaining 6 microarray data sets. Figure 6.5 and Table 6.6 show the validation results when considering KEGG, OPHID and microRNA.org as secondary data sources. The validation data sets confirm the training results: combining the three individual models significantly improved classification in 4 out of 6 data sets, while no significant improvement was obtained in the 2 remaining sets. Overall, the best individual secondary data source, the mean fixed rule and the weighted trained rule outperformed the baseline model with a p-value of 0.0313, 0.0313 and 0.0313, respectively. These higher p-values compared to the results on the training data are due to the lower number of validation data sets. When applying the Wilcoxon signed ranks test to all 16 considered data sets, the p-values decreased to 0.0004, 0.0005 and 0.001, respectively. We can therefore conclude, based on the training and validation data sets, that averaging the predictions of the three classifiers performs best. Weighting the predictions according to the training AUC values did not provide additional value. The benefit of incorporating secondary data sources was largest for data sets T1, V3 and V4 due to weaker experimental data caused by the limited number of samples.

Table 6.5: Influence of secondary data sources and classifier combination on the performance for 10 training microarray data sets.

	Model (combination)	Source/rule$^{\rho}$	Mean AUC (std)*	p-value$^{\circ}$	-log(p)
T1					
	baseline		0.782 (0.113)		
	best individual	KEGG	0.840 (0.100)	1.51e-25	24.82
	best fixed rule	mean	0.813 (0.105)	2.00e-17	16.70
	best trained rule	EXH	0.832 (0.104)	1.28e-20	19.89
	best advanced rule	LR	0.838 (0.105)	3.39e-12	11.47
T2					
	baseline		0.803 (0.081)		
	best individual	OPHID	0.805 (0.078)	0.199	0.70
	best fixed rule	mean	0.808 (0.081)	0.0113	1.95
	best trained rule	W	0.808 (0.081)	0.0114	1.94
	best advanced rule	NB	0.806 (0.081)	0.125	0.90
T3					
	baseline		0.571 (0.073)		
	best individual	OPHID	0.586 (0.078)	6.11e-5	4.21
	best fixed rule	mean	0.576 (0.076)	0.0121	1.92
	best trained rule	W	0.575 (0.077)	0.089	1.05
	best advanced rule	LDA	0.550 (0.085)	0.999	0.00
T4					
	baseline		0.748 (0.088)		
	best individual	KEGG	0.762 (0.085)	8.17e-7	6.09
	best fixed rule	median	0.765 (0.086)	7.30e-15	14.14
	best trained rule	W	0.764 (0.085)	1.87e-15	14.73
	best advanced rule	NB	0.764 (0.087)	3.40e-15	14.47
T5					
	baseline		0.727 (0.089)		
	best individual	KEGG	0.736 (0.091)	0.000437	3.36
	best fixed rule	mean	0.737 (0.090)	6.02e-6	5.22
	best trained rule	W	0.737 (0.090)	8.77e-6	5.06
	best advanced rule	NB	0.734 (0.090)	0.000644	3.19
T6					
	baseline		0.645 (0.078)		
	best individual	OPHID	0.653 (0.077)	9.73e-9	8.01
	best fixed rule	mean	0.650 (0.078)	2.74e-7	6.56
	best trained rule	W	0.650 (0.078)	3.27e-7	6.48
	best advanced rule	NB	0.647 (0.077)	0.0747	1.13

Continued on next page

Table 6.5 – continued from previous page

Data set	Model (combination)	Source/rule$^{\rho}$	Mean AUC (std)*	p-value°	-log(p)
T7					
	baseline		0.929 (0.056)		
	best individual	KEGG	0.942 (0.048)	6.02e-10	9.22
	best fixed rule	mean	0.937 (0.050)	2.27e-11	10.64
	best trained rule	W	0.937 (0.050)	3.45e-11	10.46
	best advanced rule	NB	0.937 (0.050)	1.36e-10	9.87
T8					
	baseline		0.584 (0.096)		
	best individual	microRNA.org	0.587 (0.096)	0.156	0.81
	best fixed rule	median	0.581 (0.096)	0.785	0.10
	best trained rule	W	0.579 (0.093)	0.943	0.03
	best advanced rule	NB	0.565 (0.096)	0.999	0.00
T9					
	baseline		0.573 (0.110)		
	best individual	OPHID	0.592 (0.117)	0.0151	1.82
	best fixed rule	mean	0.594 (0.113)	0.00842	2.07
	best trained rule	W	0.594 (0.113)	0.00787	2.10
	best advanced rule	NB	0.566 (0.117)	0.797	0.10
T10					
	baseline		0.680 (0.070)		
	best individual	microRNA.org	0.691 (0.065)	0.0176	1.75
	best fixed rule	mean	0.688 (0.065)	0.0596	1.23
	best trained rule	SW	0.689 (0.065)	0.0419	1.38
	best advanced rule	NB	0.688 (0.066)	0.0586	1.23

$^{\rho}$Secondary data source or combination rule with the best performance (W = weighting according to the 5-fold training AUC values; SW = weighting with scaled 5-fold training AUC values; EXH = exhaustive search among all possible combinations of 3 classifiers; NB = naïve Bayes; LR = logistic regression; LDA = linear discriminant analysis). *Mean test AUC value under best training (standard deviation). °One-sided paired-sampled t-test for the comparison with respect to the baseline model.

Table 6.6: Performance of the three selected secondary data sources and the combination rule per type that performed best on the training data (mean, AUC weighting (W) and naïve Bayes (NB)) for 6 validation microarray data sets.

	Source/rule	Mean AUC (std)	p-value°	-log(p)
V1				
		0.741 (0.099)		
	KEGG	0.739 (0.097)	0.807	0.09
	OPHID	0.734 (0.097)	0.982	0.01
	microRNA.org	0.754 (0.101)	1.18e-7	6.93
	mean	0.753 (0.096)	1.37e-7	6.86
	W	0.752 (0.096)	3.13e-7	6.50
	NB	0.750 (0.096)	3.37e-6	5.47
V2				
		0.610 (0.171)		
	KEGG	0.593 (0.154)	0.899	0.05
	OPHID	0.627 (0.146)	0.0928	1.03
	microRNA.org	0.610 (0.151)	0.509	0.29
	mean	0.623 (0.147)	0.160	0.80
	W	0.623 (0.148)	0.157	0.80
	NB	0.595 (0.157)	0.864	0.06
V3				
		0.734 (0.225)		
	KEGG	0.770 (0.220)	2.27e-10	9.64
	OPHID	0.723 (0.230)	0.979	0.01
	microRNA.org	0.749 (0.224)	4.40e-4	3.36
	mean	0.753 (0.225)	4.91e-6	5.31
	W	0.753 (0.225)	5.10e-6	5.29
	NB	0.737 (0.234)	0.380	0.42
V4				
		0.652 (0.189)		
	KEGG	0.688 (0.183)	5.11e-9	8.29
	OPHID	0.630 (0.195)	0.999	0.00
	microRNA.org	0.702 (0.180)	4.45e-17	16.35
	mean	0.682 (0.182)	4.55e-9	8.34
	W	0.682 (0.182)	6.12e-9	8.21
	NB	0.661 (0.182)	0.112	0.95
V5				
		0.667 (0.186)		
	KEGG	0.679 (0.180)	0.0128	1.89
	OPHID	0.699 (0.176)	1.75e-6	5.76

Continued on next page

Table 6.6 – continued from previous page

	Source/rule	Mean AUC (std)	p-value°	-log(p)
	microRNA.org	0.669 (0.187)	0.338	0.47
	mean	0.691 (0.180)	2.95e-7	6.53
	W	0.692 (0.179)	1.55e-7	6.81
	NB	0.678 (0.187)	0.0428	1.37
V6				
		0.604 (0.110)		
	KEGG	0.603 (0.109)	0.656	0.18
	OPHID	0.618 (0.107)	0.00143	2.85
	microRNA.org	0.611 (0.109)	0.00398	2.40
	mean	0.607 (0.106)	0.196	0.71
	W	0.607 (0.106)	0.185	0.73
	NB	0.596 (0.114)	0.949	0.02

°One-sided paired-sampled t-test for the comparison with respect to the baseline model.

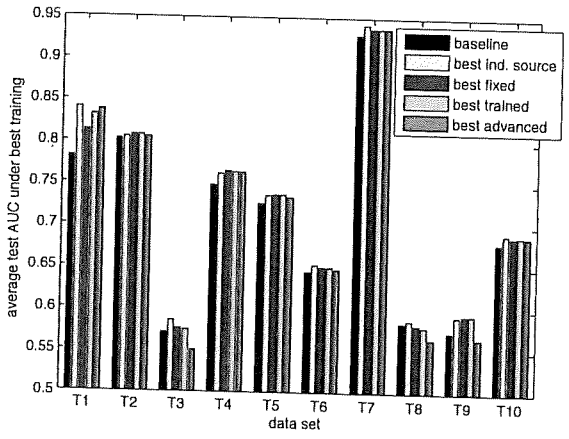


Figure 6.3: *Overall training results:* Influence of secondary data sources and classifier combination on classification performance. The average test AUC values under best training for 10 training microarray data sets, referred to as T1 to T10 are shown. Per training set, five bars are shown: 1) mean AUC of the baseline model (black); 2) mean AUC for the best individual secondary data source (KEGG, OPHID or microRNA.org) (light gray); 3) mean AUC for the best fixed combination rule (dark gray); 4) mean AUC for the best trained combination rule (light gray); 5) mean AUC for the best advanced model (dark gray). The secondary data source and combination rules that performed best and the p-values for the comparisons per training set are shown in Table 6.5.

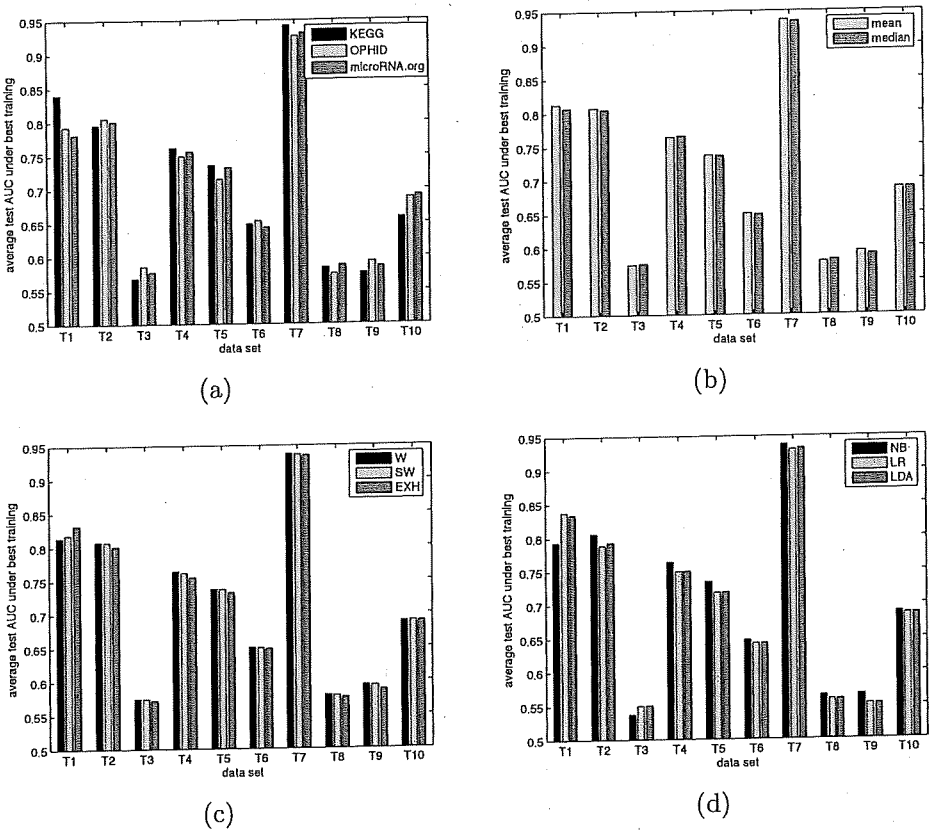


Figure 6.4: *Training results per type of combination rule:* Average test AUC values under best training for classifiers or combined classifiers, each built on one of the outstanding secondary data sources KEGG, OPHID or microRNA.org for 10 training microarray data sets. **(a)** mean AUC values for the individual classifiers, built on KEGG (black), OPHID (light gray) or microRNA.org (dark gray); **(b)** mean AUC values for two fixed rules: mean (light gray) and median (dark gray); **(c)** mean AUC values for three trained rules: weighted with the 5-fold training AUC values (W) (black), weighted with the 5-fold training AUC values scaled to]0,1] (SW) (light gray) and exhaustive search among all possible combinations of 3 classifiers (EXH) (dark gray); **(d)** mean AUC values for three more advanced models: naïve Bayes (NB) (black), logistic regression (LR) (light gray) and linear discriminant analysis (LDA) (dark gray).

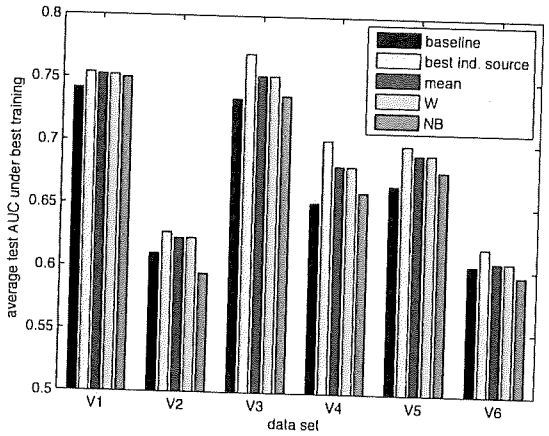


Figure 6.5: *Validation results for the optimal combination rules:* Validation of the three selected secondary data sources and the best performing combination rules on six new microarray data sets, referred to as V1 to V6. Per validation data set, five bars with the average test AUC under best training are shown: 1) mean AUC of the baseline model (black); 2) mean AUC for the best individual secondary data source (KEGG, OPHID or microRNA.org) (light gray); 3) mean AUC for the fixed mean rule (dark gray); 4) mean AUC for the weighted combination rule (light gray); 5) mean AUC for the naïve Bayes rule (dark gray). A numerical overview is given in Table 6.6.

6.5 Conclusion

Improved outcome prediction and classification decision making have been made possible by incorporating the human interactome in a wLS-SVM model, chosen among a large set of possible methods in which prior information can be incorporated in the presented manner. Interactome data from secondary data sources were encoded in a graph-based way and used in similarity matrices for patient classification. Ten microarray data sets were randomly selected for training. For the majority of these data sets, three secondary data sources increased performance with respect to the baseline model based on microarray data only, being KEGG, OPHID and microRNA.org. The 15 secondary databases considered in this study could therefore be reduced to three gene-related information sources that are relevant across multiple cancer types with regard to the considered class of models. These sources also showed a good performance on 6 validation microarray data sets. With only three models to combine, it could be shown on all data sets that averaging the predictions of the individual models suffices. However, when sample sizes increase, more reliable estimates of the weights may be obtained so that optimization of the weights becomes beneficial. Combination rules that require additional training are therefore only expected to gain in importance when

the sample size in microarray studies would increase.

The three selected sources KEGG, OPHID and microRNA.org are the most complete databases for their type of information. With the KEGG database, information is incorporated about groups of genes involved in the same metabolic pathways, pathways related to genetic information processing, regulatory pathways or pathways active in human diseases and drug development. OPHID is a database on predicted interactions between human proteins combining PPIs obtained from specialist literature, high-throughput experiments, evolutionary conservation and other databases such as BIND, HPRD and MINT. In the microRNA.org database, predicted miRNA target sites are filtered according to evolutionary conservation. An improved classification performance by including the microRNA.org network confirms the importance of miRNAs in cancer and their impact on target genes.

The extra layer for the combination of classifiers is essential. For verification, the largest network per type of gene-related information, being KEGG, REACTOME, OPHID, STRING, PROSITE and microRNA.org, were pooled with an early integration approach (that is, the edge weights were added) before a wLS-SVM model was built. However, in all cases our 2-layer approach outperformed models built on a union of secondary data sources. This emphasizes the specific edge interpretation for each type of interaction and makes the construction of individual models per secondary data source essential before combining them at a second level.

The idea of representing prior knowledge as a graph before incorporating it into a classification or regression framework to improve performance is not restricted to kernel methods. We have extended this idea to a general regression framework [298]. For many diseases, markers coincide with entire groups of associated variables, thus enforcing the construction of predictive models in terms of these *macro variables*. Whereas classical statistical approaches treat genes as independent and equally important variables, prior knowledge of known interaction patterns has previously been incorporated in a general regression framework with network-constrained regularization (NET), as depicted in the introduction of this chapter [205]. An inaccurate graph with a considerable amount of misspecified associations, however, can decrease the performance of the model constructed with the aid of the graph prior. As multiple graphs are available, we explored how to combine these priors in an improved regression framework (multiple-NET) [298]. For the microarray data set T1 on ovarian cancer, the graphs derived from the sources miRNAmap and microRNA.org were considered as priors. The accuracy was measured by repeating 5-fold CV 100 times. Multiple-NET with an averaged CV AUC of 0.82 outperformed the NET approach applied with each of the priors individually (both 0.80) and simpler approaches only based on microarray data (lasso 0.76; elastic net 0.78).

In conclusion, we showed that it is possible to incorporate prior information from secondary data sources in the form of the human interactome in a support vector machine and a general regression framework. Any type of gene-related information can be considered and the matrices derived from each graph are representable in many different ways. Moreover, no part of the data needs to be discarded. All genes are considered, also the ones that have not been thoroughly investigated or for which no annotation is available. Our results show that for the considered microarray technologies, cancer types and types of outcome, integrating interactome data improves classification of cancer outcome based on microarray data. This integration of prior information in an SVM model based on gene expression may benefit investigation of biological functionality.

Chapter 7

Conclusions and Future Research

7.1 Conclusions

The growing amount of genome-wide data combined with factors such as time, cost and personalized treatment is complicating clinical decision making. The aim of this dissertation is therefore to use advanced mathematical models such as the support vector machine to aid clinical decision support. While information arising from clinicopathology (for example, patient history, laboratory analysis or ultrasound parameters) is not accurate enough to reliably predict patient prognoses, microarray technology takes into account the fundamental mechanisms underlying carcinogenesis on a molecular level. Although many breakthroughs have been made and the first gene expression profile signatures are being validated in clinical trials, the promise of individualized therapy has not yet been fulfilled. The transcriptome measured with microarray technology does not completely reflect the biology underlying a disease. New technologies profiling other levels in the central dogma of molecular biology are emerging. Studying these -omics layers such as the genome and proteome can add important complementary information to predict the outcome in cancer patients. We therefore focused not only on the integration of multiple data sets gathered from the same patients to improve clinical decision support, but also on the incorporation of prior knowledge about the human interactome, and finally on an improved modeling of clinical data, which should not be disregarded as routinely obtained in daily practice.

Although several research groups have already proposed the idea of data integration, they have only investigated the integration of clinical and microarray

data, analyzed two -omics data sources sequentially, or merged data sets at the database level. Our definition for the combination of heterogeneous biological data differs from what has been described in specialist literature. We integrated multiple layers of experimental data into a machine learning-based model for the development of more homogeneous classifiers, with the most relevant features selected from all available data sources. For this purpose, we presented a kernel-based integration framework within the patient domain in Chapter 3. We first showed that for breast cancer clinical data, which are the basis of day-to-day clinical decision support, can be combined efficiently with microarray data to improve patient tailored therapy. Subsequently, applying our algorithm to transcriptomics and proteomics data for rectal cancer and genomics with transcriptomics data for prostate cancer showed that data from different layers in the central dogma are partly complementary. Combining them improved performance of decision support in cancer upon models based on data with one type of biological information or on a combination of less data sets than available.

While most attention in specialist literature has been paid to transcriptome-based classification, we wanted our integration framework to also be applicable to more complex data types from which prognostic biomarkers can be extracted. We therefore developed a workflow for both array CGH and mass spectrometry data, to convert data into features that are incorporable in our classification framework. For mass spectrometry data, a workflow for the conversion of proteomic spectra into a list of peaks that are differential in two different groups of patients was proposed in [46], whilst our focus in Chapter 5 was on array CGH data. Per group of patients, we constructed a recurrent HMM in which a cohort of samples instead of an individual sample is modeled to increase statistical power by sharing information on copy number variations. The differential regions, defined as chromosomal regions that are gained/lost in one group of patients while not being gained/lost in the other group, were reduced to a subset of features for classification. The combination of recurrent HMMs for the detection of copy number alterations with support vector machines offers a novel methodological approach for classification based on copy number alterations. Additionally, this approach limits the chromosomal regions that are necessary to classify patients according to cancer subtype. Knowledge on CNVs that are different in specific groups of patients may help in understanding the tumorigenesis of these cancers. These two strategies, proposed in [46] and Chapter 5, allow spectra and CNV data to be included as an additional data source in the proposed kernel-based integration framework for clinical decision support.

As microarray technology was the most advanced in the past decade, the majority of studies only gathered transcriptomics data. This data type, however, has been subject to criticisms and concerns. To obtain more stable gene signatures in microarray studies, to improve and robustify diagnosis and prognosis and to decipher the biological deregulated processes in which these genes are involved, in

Chapter 6 we incorporated prior biological knowledge from databases with various, distinct gene information in a classification model, with gene expression levels as predictors and diagnostic or prognostic categories as targets. Whereas with kernel methods similarity between patients is traditionally calculated with a similarity measure based on the patients' gene expression profiles, interactome data from any database can be used to improve the way patient similarities are calculated and thus also to improve classification performance, by exploiting relations between genes with similar functions but active in alternative pathways. Of all considered databases, three databases (one on metabolic pathway information, one on protein-protein interactions and one on miRNA-gene targeting) significantly improved prediction of a wide range of cancer-related outcomes in 10 microarray data sets used for training. As the prior knowledge database that leads to the largest increase in classification performance depends on the specific microarray data set or outcome, and as inclusion of irrelevant prior knowledge may worsen classification, multiple approaches were considered to combine the three corresponding models. Averaging the predictions of these models performed best and was significantly better than the model only based on microarray data. Confirmation on 6 other microarray data sets illustrated an improved classification with prior biological knowledge for the investigated cancer types and microarray platforms.

In addition to all studied -omics data types, clinical data are available in the majority of studies in cancer research. To fully exploit this information, appropriate modeling is required. In Chapter 4, we proposed a linear additive kernel function that distinguishes between continuous, ordinal and nominal variables. Both this alternative and its nonlinear version, obtained in analogy with expanding the linear to the polynomial kernel function, were considered. Support vector machines based on our kernel function significantly outperformed the widely used linear and polynomial kernel function when tested on four clinical data sets and for the combination of clinical and microarray data. Not discussed in Chapter 4 but also important for an improved modeling of clinical data is Functional Linear Discriminant Analysis (FLDA). In the past longitudinal data were wrongly analyzed with cross-sectional approaches by assuming independence between data points from the same patient or by disregarding data points to keep only one measurement per patient, thereby wasting data and weakening statistical power. To overcome this problem of co-dependence, we introduced FLDA for the analysis of longitudinal data in the field of biomedicine [34] [75] [76]. FLDA can be applied to decipher the relationship between any two variables. We discriminated normal and abnormal growth in the first trimester of pregnancy and were able to better predict miscarriage with FLDA based on growth than with a single observation of the embryo's size [34]. We also established the relationship of umbilical artery and middle cerebral artery Doppler impedance with abdominal circumference in babies with severe early onset fetal growth restriction [320], and showed an association between first trimester PAPP-A level and birthweight [150].

7.2 Future Research

Despite our efforts to bridge the gap between clinicians and engineers, a closer collaboration between both groups is still essential for breakthroughs and translational medicine to become true. An overview of challenges due to fast evolving technologies and challenges from a methodological point of view is given, followed by a view on the future of clinical decision support.

7.2.1 Medical and Technological Trends

Currently, it is unknown which technology and thus which level of molecular biology is most relevant for outcome prediction. To assess the most relevant omics platform, studies are needed in which multiple technologies are applied to the same tumor samples in sufficiently large number. Although this will substantially increase costs, it is a necessary investment to move away from costly empirical models towards improving cancer diagnosis and tailoring treatment for individual patients. Furthermore, by limiting the amount of biomarkers, small arrays or qPCR can be used for diagnostic testing.

However, there is a shift away from the technologies studied in this dissertation due to growing maturity of other technologies. DNA sequencing technology with its increased power and steadily falling costs is revolutionizing the field of personal genomics, with new opportunities for software and algorithm developments and greater emphasis on supercomputing infrastructure [18] [177] [201] [207] [295]. Sequencing has the capability to replace microarray technology as several earlier technical limitations have been circumvented (that is, no prior sequence knowledge, no cross-hybridization, digitized quantization and reproducibility). Although it is currently too expensive for many labs to procure sequencing equipment and computing infrastructure, it is expected that biologists are willing to spend more to get better data. Thus, in some areas of life sciences, sequencing does put microarrays at risk. On the other hand, as our genetic understanding of diseases steadily grows, microarrays will become increasingly useful for tailor-made, small-market arrays in clinical diagnostics and for specialized applications, such as exon sequencing and the study of epigenetic modifications. Other improvements at the technological level can also be expected. Currently, for example, technical limitations hamper progress in examining genome-wide epigenetic silencing by methylation. Improvements are required to obtain insights into how these epigenetic modifications affect development and disease.

Biomarkers that are selected for their prognostic capability are not necessarily good targets for treatment [109]. By only focusing on prognostic markers that are selected from a single data set, important key players in cancer may be missed and knowledge about cancer pathology remains incomplete. The development

of more sensitive, reliable, non-invasive and cost-effective methods for the early detection of cancer is a priority in translational cancer research. Aberrant DNA methylation and histone acetylation have become attractive as targets for cancer epigenetic therapy. The intrinsic reversibility of epigenetic modifications allows genes that are epigenetically silenced in cancer to be re-activated. This represents an opportunity to combine epigenetic agents with chemotherapy at lower dosages. These novel strategies will enhance the efficacy of therapy, while at the same time lowering toxicity and enhancing patient response [158] [305].

Finally, as has been shown, a "one drug fits all" approach is not appropriate because subtypes within each cancer type do exist. Each subtype should therefore be administered a different drug. With technologies becoming more mature and validated, bench-to-bedside clinical applications rapidly emerge. A trend is seen from a disease-centric approach, in which all patients within a subtype are administered the same dose, to a patient-centric approach, taking into account patient-specific genetic and epigenetic abnormalities with differences in drug metabolism. This requires trials to be tailored to relevant molecular subtypes. A special effort is also required to collect patient data and tumor material in biobanks according to well-defined standard operating procedures, in order to obtain inter-lab comparability and high-quality data. [309].

7.2.2 Methodological Challenges

Besides the support vector machine that was used in this dissertation to improve clinical decision support in cancer, Bayesian networks have been investigated for data integration as well [137]. Although not tuned for decision support and requiring data from many patients, Bayesian networks allow learning probabilistic interactions between the most important features within one level of the genome or between different levels when considering multiple genome-wide data sources. In the future, our kernel-based integration method should therefore be compared with the Bayesian integration framework. Moreover, as these frameworks are complementary, an ensemble approach can be applied. More accurate classifiers are likely to be obtained, not only by combining different data types but also individual decisions of multiple classifiers. In this way, the advantages of both methods are exploited.

Furthermore, the optimization algorithm, especially for the choice of the weights, can be improved. In Chapter 3, an average combination of kernel matrices was considered. Learning from all available data sets, however, is not always a good idea. When some of the available data sets provide much more information than others, the latter sets may contribute noise that disrupts learning. In case the amount of available data sets further increases, the choice of the weights will become even more important, especially when the algorithm is applied to

problems for which the reliability of the data sources differs much or is not known *a priori*. For multiple kernel learning, we therefore thoroughly investigated the use of different norms for the error term in the objective function of the SVM and LS-SVM [378]. L_1 regularization is characterized by sparseness in the weights assigned to the kernel matrices. Dominant weights are assigned to only one or two matrices, which is useful in distinguishing relevant from irrelevant data sets. The L_2 regularization, on the other hand, leads to a non-sparse solution with the weights smoothly distributed over multiple matrices, avoiding a dominant data set to get a weight close to 1. This may be useful in case of carefully selected and complementary data sets. The approach in this dissertation is situated in between these two types of regularization. For multiple biomedical applications among which the one on rectal cancer investigated in Chapter 3 and the pure clinical data sets from Chapter 4, we showed that L_2 regularization outperformed L_1 regularization. L_2 regularization is therefore a promising approach for genome-wide data fusion applications [378].

However, the focus should not be restricted to the selection of good weights. In Chapter 3, three integration types were presented among which we chose the intermediate integration approach. Although the use of this approach is, due to its advantages, defendable for many applications, it may be possible that early or late integration perform better in some domains. Ideally, the integration type and the data sets to incorporate should be chosen before training. An important question therefore is whether a decision about the best performing approach can be extracted from the data itself before applying tools for clinical decision support, thereby fully formalizing data integration within any domain. Moreover, the amount of data is still growing. This implicates that for each cancer type, newer and more accurate data sets will become accessible on top of the widely available genomics, transcriptomics and proteomics data. This raises the question of how the usefulness and complementarity of an extra data set should be assessed, to decide on expansion of existing models.

More advanced feature selection techniques are also required. A disadvantage of univariate feature selection in biomedical applications is the simultaneous selection of highly correlated features as biomarkers, leading to cost-inefficient models for clinical diagnosis and prognosis. The domain of feature selection is likely to evolve due to the trend towards multicentric collaborations in which larger data sets are collected.

Finally in Chapter 6, we were forced to consider networks with undirected links. Current databases lack information on whether a gene is activated or repressed at transcription level, or whether proteins in a PPI transiently modify each other or form a protein complex. A systematic and detailed annotation of information about the influence of interactions on cell state is a major challenge in systems biology [167], and therefore a prerequisite for improved CDS in cancer research. Furthermore, we did not intend to exhaustively consider all possible secondary

databases. Other databases in each of the five categories (see Section 6.2.2) could have been included as well. Two categories with potential value that were not considered are drug- and literature-based databases. The database STITCH [191] contains protein-chemical associations from 630 organisms with 74,000 different chemicals, and interactions between human proteins and 2,200 drugs or drug-like molecules. Two examples of literature-derived information with links between genes that are not available in traditional pathway databases are PubGene [175] and the GeneWays Knowledge Base [276]. PubGene is a gene-to-gene co-citation network for 13,712 known human genes [175]. An automatic analysis of titles and abstracts in over 10 million MEDLINE records, of which ~20% containing one or more gene symbols, resulted in ~140,000 gene pairs with 7,512 genes involved. GeneWays, on the other hand, contains information on molecular interactions, extracted from the text of journal articles [276].

7.2.3 The Future of Clinical Decision Support

As depicted in Chapter 1, efforts are highly necessary to improve the acceptability of clinical decision support. Acceptability, however, depends on the type of CDS system. Firstly, models or assays can serve decision making; secondly, sets of biomarkers and chromosomal regions can be investigated to unravel the underlying tumorigenesis and to detect deregulated processes; and finally, support tools can aid visualization. In this dissertation, results in all three categories were obtained.

Concerning the approaches from the first category that were proposed in this dissertation, much research is required before they will reach the clinic. Although data integration and prior knowledge in combination with machine learning techniques contribute to improved decision support, the requirement for multiple high-throughput technologies and the use of a black-box model make clinical acceptability difficult. On the other hand, models from the last category are easily accepted without need for FDA-approval. The FLDA technique, for example, warns the gynecologist earlier in case of deviating growth during pregnancy than would be possible based on the embryo's size. Hence, the less interpretable the model, the higher the threshold for clinicians to accept it, and the more difficult to get the model approved by FDA. FDA distinguishes various types of CDS systems according to the level of control required to assure safety and effectiveness, the complexity of the test and its multivariate characteristic.

Despite clinical trials in which a gene signature is tested with respect to a clinicopathologic criterion, a trend is nowadays seen towards research proposals on the investigation of multiple types of data, gathered from the same set of patients. Although obtaining interpretable models is even more challenging with the increase in complexity and number of -omics data, the identification of heterogeneous sets of markers is a state-of-the-art research topic. With this recent

trend, broader data sets will become available to further improve our methods in clinical interpretability and acceptability and from an algorithmic point of view. Moreover, the multi-modal data sets should be obtained from a sufficient number of patients to increase the reliability of a decision support tool or a set of biomarkers, and to have enough power to prove a clinically useful increase in performance with respect to standard tests or clinical indices.

As the IT dimension is very important to fill the translational gap between research and clinical practice and as the methods proposed in this dissertation are proof-of-concept, further research on the software component is required to develop advanced software tools oriented towards clinical applicability. Once these techniques find their way into the clinical practice, multiple standard clinical tests can be replaced by a more efficient, multi-modal assay to counteract medical overconsumption (for example, by reducing the fraction of cancer patients that are administered chemotherapy without worsening disease-free survival). In this way, CDS systems will impact our medical care system by improving cost-effectiveness. Such assays should also be reimbursed by the health care providers, in order to improve quality of health on a population-wide basis.

Finally, before a CDS system can be operational, legal and ethical issues need to be addressed. Another issue, which becomes more important partly due to the realization of the $1,000 genome, is that of privacy and confidentiality. Although an individual's privacy needs to be guaranteed, an electronic patient record would substantially attribute to the success of CDS, by making large databases with clinical and genomic information available for research and decision support development. In conclusion, the only solution to improved health care is the convergence of distinct domains, as many challenges in different areas need to be tackled before the realization of multivariate CDS systems can become a fact.

Appendix A

ROC Curves

In this appendix, we show the ROC curves of the wLS-SVM models for those combinations of data sets shown in Table 3.2 for which only one technology or a sample at one time point is required, and for all models shown in Table 3.4. Figures A.1 to A.3 show the ROC curves for the prediction of WHEELER, pN-STAGE and CRM in rectal cancer. For prostate cancer, the ROC curves for the prediction of GRADE, STAGE, METASTASIS and RECURRENCE are shown in Figures A.4 to A.7.

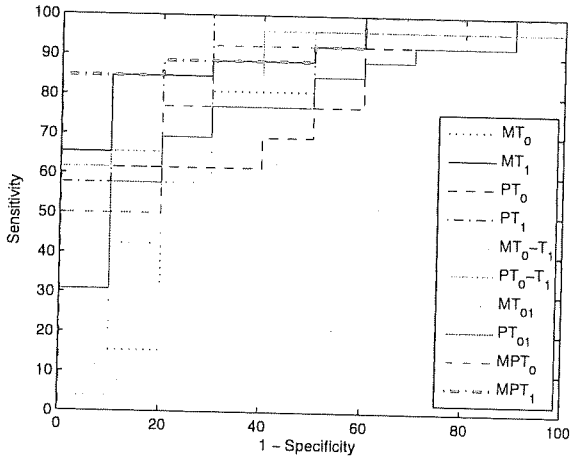


Figure A.1: *WHEELER* – MT_0 (4 genes, black dotted), MT_1 (29 genes, black solid), PT_0 (35 proteins, black dashed), PT_1 (11 proteins, black dashdotted), $MT_0 - T_1$ (32 genes, light gray dotted), $PT_0 - T_1$ (5 proteins, light gray solid), MT_{01} (3 genes, dark gray dotted), PT_{01} (21 proteins, dark gray solid), MPT_0 (3 genes, 35 proteins, dark gray dashed), and MPT_1 (25 genes, 12 proteins, dark gray dashdotted)

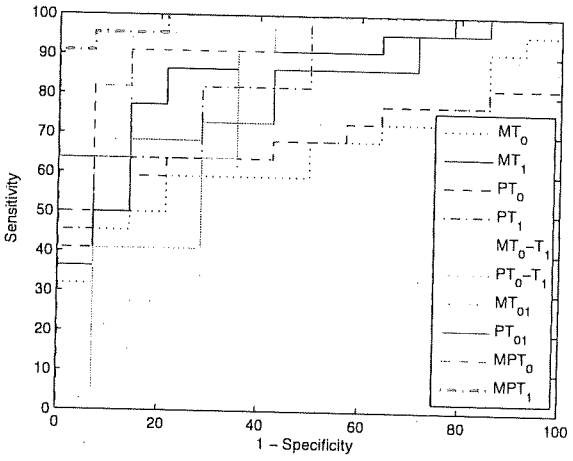


Figure A.2: *pN-STAGE* – MT_0 (25 genes, black dotted), MT_1 (22 genes, black solid), PT_0 (2 proteins, black dashed), PT_1 (12 proteins, black dashdotted), $MT_0 - T_1$ (4 genes, light gray dotted), $PT_0 - T_1$ (9 proteins, light gray solid), MT_{01} (24 genes, dark gray dotted), PT_{01} (34 proteins, dark gray solid), MPT_0 (27 genes, 27 proteins, dark gray dashed), and MPT_1 (21 genes, 14 proteins, dark gray dashdotted)

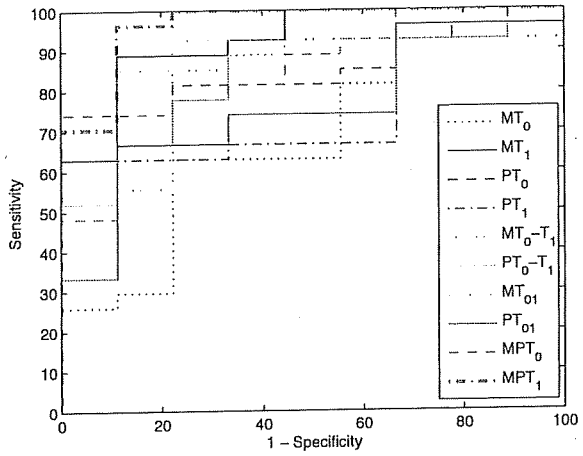


Figure A.3: $CRM - MT_0$ (33 genes, black dotted), MT_1 (9 genes, black solid), PT_0 (34 proteins, black dashed), PT_1 (34 proteins, black dashdotted), $MT_0 - T_1$ (6 genes, light gray dotted), $PT_0 - T_1$ (2 proteins, light gray solid), MT_{01} (16 genes, dark gray dotted), PT_{01} (3 proteins, dark gray solid), MPT_0 (7 genes, 27 proteins, dark gray dashed), and MPT_1 (7 genes, 33 proteins, dark gray dashdotted)

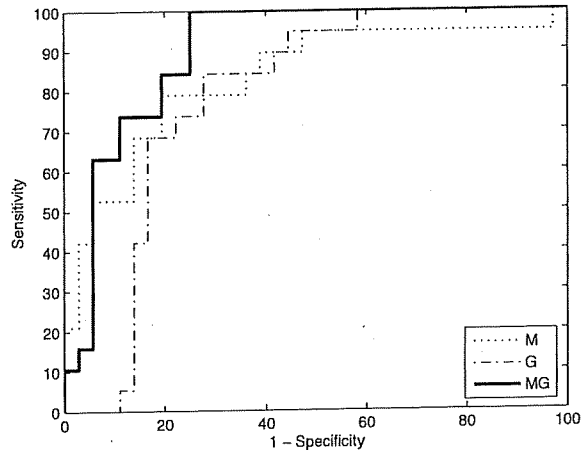


Figure A.4: $GRADE - M$ (24 genes, dotted), G (8 CNVs, dashdotted), and MG (6 genes, 8 CNV, solid)

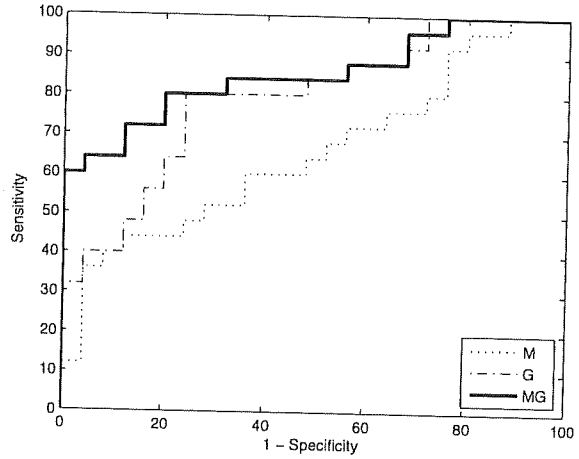


Figure A.5: *STAGE* – M (18 genes, dotted), G (32 CNVs, dashdotted), and MG (42 genes, 22 CNV, solid)

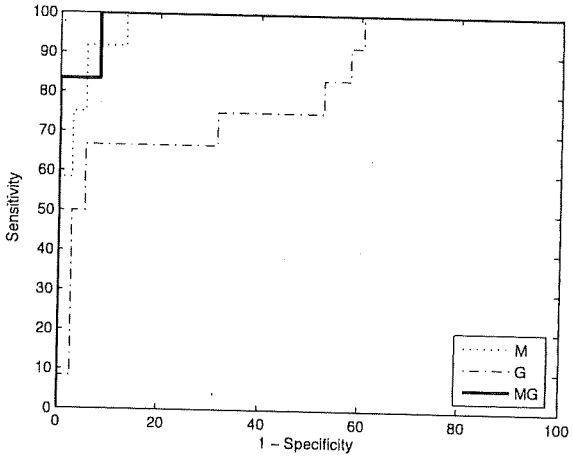


Figure A.6: *METASTASIS* – M (18 genes, dotted), G (12 CNVs, dashdotted), and MG (18 genes, 3 CNV, solid)

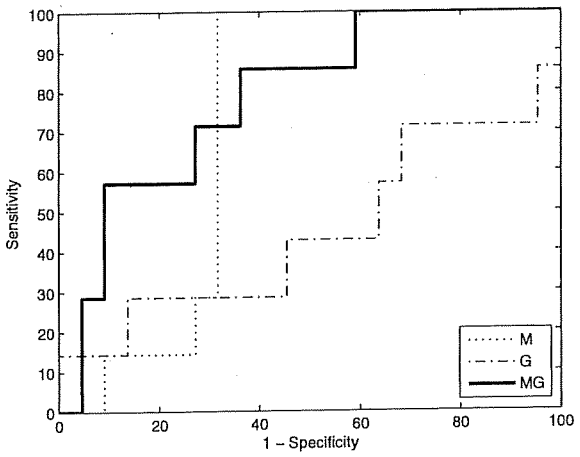


Figure A.7: *RECURRENCE* – M (24 genes, dotted), G (26 CNVs, dashdotted), and MG (32 genes, 2 CNV, solid)

Appendix B

Generalizability Results

To get an idea on the variability of the reported LOO performances as well as to compare the LOO results with test results, we performed 20 randomizations for wLS-SVM models based on different combinations of data sets for the prediction of response in rectal and prostate cancer. Since the randomizations are not disjoint, the test results can still be too optimistic, but already give a more reliable idea on generalization ability. The size of the feature subsets we considered ranged from 1 to 35 in steps of 5. In each randomization, all data were randomly split, stratified to outcome, into a training set containing $5/6^{th}$ of samples and a test set containing the other $1/6^{th}$. For training, LOO was used to optimize regularization parameter γ. The wLS-SVM model with the optimal γ was validated on the test set, using those features that were selected most often in all LOO iterations. For each outcome, this strategy resulted in 20 internal LOO and 20 test performances for each considered combination of number of features.

This appendix shows for the different combinations of data sets the internal LOO performances (AUC) in subplot 1 for the wLS-SVM models with optimal number of features. These boxplots are compared with their respective test performances in subplot 2. Figures B.1 and B.2 show the comparisons for pN-STAGE and CRM in rectal cancer, while the results for STAGE, METASTASIS and RECURRENCE in prostate cancer are presented in Figures B.3 to B.5.

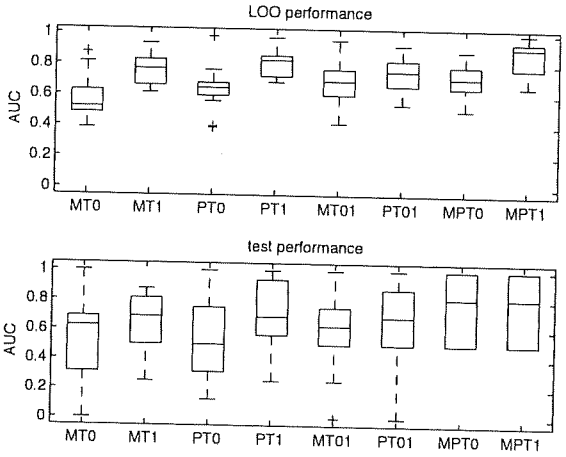


Figure B.1: *pN-STAGE* – comparison of the internal LOO and test performances for MT_0 (30 genes), MT_1 (35 genes), PT_0 (30 proteins), PT_1 (25 proteins), MT_{01} (35 genes), PT_{01} (25 proteins), MPT_0 (25 genes, 25 proteins), and MPT_1 (30 genes, 20 proteins)

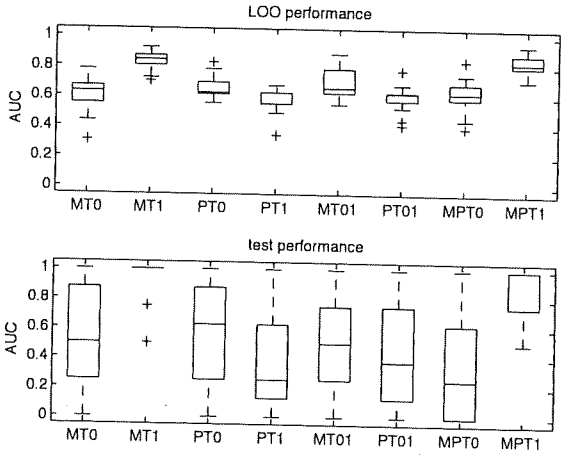


Figure B.2: *CRM* – comparison of the internal LOO and test performances for MT_0 (30 genes), MT_1 (25 genes), PT_0 (30 proteins), PT_1 (30 proteins), MT_{01} (15 genes), PT_{01} (15 proteins), MPT_0 (20 genes, 10 proteins), and MPT_1 (30 genes, 1 protein)

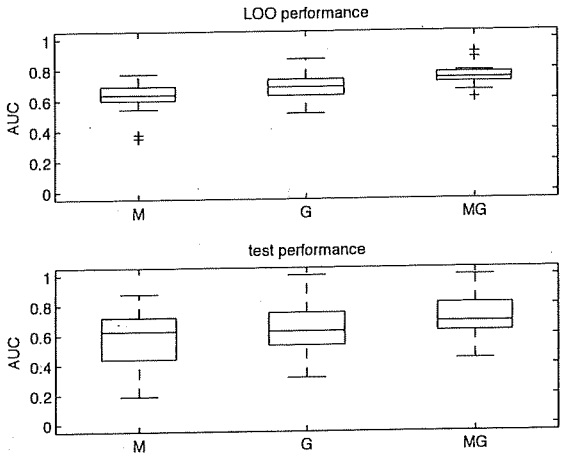


Figure B.3: *STAGE* – comparison of the internal LOO and test performances for M (35 genes), G (35 CNVs), and MG (30 genes, 35 CNVs)

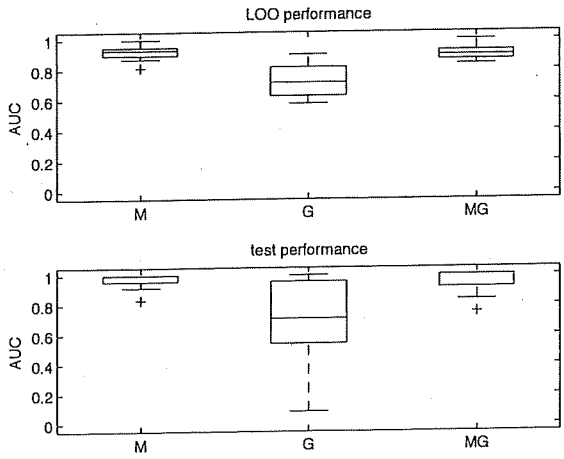


Figure B.4: *METASTASIS* – comparison of the internal LOO and test performances for M (20 genes), G (35 CNVs), and MG (25 genes, 1 CNV)

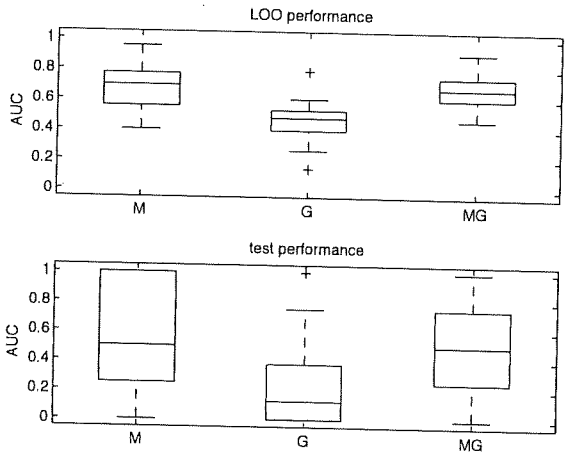


Figure B.5: *RECURRENCE* – comparison of the internal LOO and test performances for M (25 genes), G (20 CNVs), and MG (25 genes, 1 CNV)

Bibliography

[1] I J Adam, M O Mohamdee, I G Martin, N Scott, P J Finan, D Johnston, M F Dixon, and P Quirke. Role of circumferential margin involvement in the local recurrence of rectal cancer. *Lancet*, 344:707–711, 1994.

[2] M Adjouadi, N Zong, and M Ayala. Multidimensional pattern recognition and classification of white blood cells using support vector machines. *Part Part Syst Charact*, 22:107–118, 2005.

[3] R Aebersold and M Mann. Mass spectrometry-based proteomics. *Nature*, 422(6928):198–207, 2003.

[4] D H Ahn, S C Crawley, R Hokari, S Kato, S C Yang, J D Li, and Y S Kim. TNF-alpha activates MUC2 transcription via NF-kappaB but inhibits via JNK activation. *Cell Physiol Biochem*, 15(1-4):29–40, 2005.

[5] K Al-Kuraya, P Schraml, J Torhorst, C Tapia, B Zaharieva, H Novotny, H Spichtin, R Maurer, M Mirlacher, O Köchli, M Zuber, H Dieterich, F Mross, K Wilber, R Simon, and G Sauter. Prognostic relevance of gene amplifications and coamplifications in breast cancer. *Cancer Res*, 64:8534–8540, 2004.

[6] B Alberts, D Bray, K Hopkin, A Johnson, J Lewis, M Raff, K Roberts, and P Walter. *Essential cell biology*. Garland Science, New York, 2004.

[7] A Alon, N Barkai, D A Notterman, K Gish, S Ybarra, D Mack, and A J Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc Natl Acad Sci U S A*, 96:6745–6750, 1999.

[8] C Alzate. *Support Vector Methods for Unsupervised Learning*. PhD thesis, Katholieke Universiteit Leuven, Department of Electrical Engineering, 2009.

[9] C Ambroise and G J McLachlan. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc Natl Acad Sci U S A*, 99(10):3389–3402, 2002.

[10] B Aranda, P Achuthan, Y Alam-Faruque, I Armean, A Bridge, C Derow, M Feuermann, A T Ghanbarian, S Kerrien, J Khadake, J Kerssemakers, C Leroy, M Menden, M Michaut, L Montecchi-Palazzi, S N Neuhauser, S Orchard, V Perreau, B Roechert, K van Eijk, and H Hermjakob. The IntAct molecular interaction database in 2010. *Nucleic Acids Res*, 38:D525–D531, 2010.

[11] M Ashburner, C A Ball, J A Blake, D Botstein, H Butler, J M Cherry, A P Davis, K Dolinski, S S Dwight, J T Eppig, M A Harris, D P Hill, L Issel-Tarver, A Kasarskis, S Lewis, J C Matese, J E Richardson, M Ringwald, G M Rubin, and G Sherlock. Gene Ontology: tool for the unification of biology. *Nat Genet*, 25:25–29, 2000.

[12] G Bader, D Betel, and C Hogue. BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res*, 31:248–250, 2003.

[13] K A Baggerly and K R Coombes. Deriving chemosensitivity from cell lines: forensic bioinformatics and reproducible research in high-throughput biology. Posted at www.ndns.nl/static/files/sls/presentations/Baggerly-AnnalsAppliedStats.pdf.

[14] K A Baggerly, J S Morris, and K R Coombes. Reproducibility of SELDI-TOF protein patterns in serum: comparing datasets from different experiments. *Bioinformatics*, 20(5):777–785, 2004.

[15] S Bandyopadhyay, Y Wang, R Zhan, S K Pai, M Watabe, M Iiizumi, E Furuta, S Mohinta, W Liu, S Hirota, S Hosobe, T Tsukada, K Miura, Y Takano, K Saito, T Commes, D Piquemal, T Hai, and K Watabe. The tumor metastasis suppressor gene Drg-1 down-regulates the expression of activating transcription factor 3 in prostate cancer. *Cancer Res*, 66(24):11983–11990, 2006.

[16] A Barabasi and Z N Oltvai. Network biology: understanding the cell's functional organization. *Nat Rev Genet*, 5:101–113, 2004.

[17] Z Barutcuoglu, R E Schapire, and O G Troyanskaya. Hierarchical multi-label prediction of gene function. *Bioinformatics*, 22(7):830–836, 2006.

[18] A Bateman and J Quackenbush. Bioinformatics for next generation sequencing. *Bioinformatics*, 25(4):429, 2009.

[19] D W Bates, G J Kuperman, S Wang, T Gandhi, A Kittler, L Volk, C Spurr, R Khorasani, M Tanasijevic, and B Middleton. Ten commandments for effective clinical decision support: making the practice of evidence-based medicine a reality. *J Am Med Inform Assoc*, 10:523–530, 2003.

[20] A Baudot, F X Real, J M G Izarzugaza, and A Valencia. From cancer genomes to cancer models: bridging the gaps. *EMBO Rep*, 10(4):359–366, 2009.

[21] Y Benjamini and Y Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Statist Soc B*, 57:289–300, 1995.

[22] D Bentley. Whole-genome re-sequencing. *Curr Opin Genet Dev*, 16:545–552, 2006.

[23] A Berchuck, E S Iversen, J M Lancaster, J Pittman, J Luo, P Lee, S Murphy, H K Dressman, P G Febbo, M West, J R Nevins, and J R Marks. Patterns of gene expression that characterize long-term survival in advanced stage serous ovarian cancers. *Clin Cancer Res*, 11:3686–3696, 2005.

[24] D Betel, M Wilson, A Gabow, D S Marks, and C Sander. The microRNA.org resource: targets and expression. *Nucleic Acids Res*, 36:D149–D153, 2008.

[25] L Bethke, E Webb, G Sellick, M Rudd, S Penegar, L Withey, M Qureshi, and R Houlston. Polymorphisms in the cytochrome P450 genes CYP1A2, CYP1B1, CYP3A4, CYP3A5, CYP11A1, CYP17A1, CYP19A1 and colorectal cancer risk. *BMC Cancer*, 7:123, 2007.

[26] V Bewick, L Cheek, and J Ball. Statistics review 13: receiver operating characteristics curves. *Critical Care*, 8(6):508–512, 2004.

[27] H Bhaskar, D C Hoyle, and S Singh. Machine learning in bioinformatics: A brief survey and recommendations for practitioners. *Comp Biol Med*, 36(10):1104–1125, 2006.

[28] B Bhatia, C J Maldonado, S Tang, D Chandra, R D Klein, D Chopra, S B Shappel, P Yang, R A Newman, and D G Tang. Subcellular localization and tumor-suppressive functions of 15-lipoxygenase 2 (15-LOX2) and its splice variants. *J Biol Chem*, 278(27):25091–25100, 2003.

[29] A H Bild, G Yao, J T Chang, Q Wang, A Potti, D Chasse, M B Joshi, D Harpole, J M Lancaster, A Berchuck, J A Jr Olson, J R Marks, H K Dressman, M West, and J R Nevins. Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature*, 439:353–357, 2006.

[30] D A Bitton, M J Okoniewski, Y Connolly, and C J Miller. Exon level integration of proteomics and microarray data. *BMC Bioinformatics*, 9:118, 2008.

[31] P Björkholm and E L L Sonnhammer. Comparative analysis and unification of domain-domain interaction networks. *Bioinformatics*, 25(22):3020–3025, 2009.

[32] H Blegen, J S Will, B M Ghadimi, H P Nash, A Zetterberg, G Auer, and T Ried. DNA amplifications and aneuploidy, high proliferative activity and impaired cell cycle control characterize breast carcinomas with poor prognosis. *Anal Cell Pathol*, 25:103–114, 2003.

[33] J F Bosset, L Collette, G Calais, L Mineur, P Maingon, L Radosevic-Jelic, A Daban, E Bardet, A Beny, and J C Ollier. Chemotherapy with preoperative radiotherapy in rectal cancer. *N Engl J Med*, 355:1114–1123, 2006.

[34] C Bottomley, A Daemen, F Mukri, A T Papageorghiou, E Kirk, A Pexsters, B De Moor, D Timmerman, and T Bourne. Assessing first trimester growth: the influence of ethnic background and maternal age. *Hum Reprod*, 24:278–283, 2009.

[35] C Bottomley, A Daemen, F Mukri, A T Papageorghiou, E Kirk, A Pexsters, B De Moor, D Timmerman, and T Bourne. Functional linear discriminant analysis: a new longitudinal approach to the assessment of embryonic growth. *Hum Reprod*, 24:284–290, 2009.

[36] D D L Bowtell. Options available - from start to finish - for obtaining expression data by microarray. *Nat Genet*, 21:25–32, 1999.

[37] A Brand, H Brand, and T Schulte in den Bäumen. The impact of genetics and genomics on public health. *Eur J Hum Genet*, 16:5–13, 2008.

[38] B J Breitkreutz, C Stark, T Reguly, L Boucher, A Breitkreutz, M Livstone, R Oughtred, D H Lackner, J Bähler, V Wood, K Dolinski, and M Tyers. The BioGRID Interaction Database: 2008 update. *Nucleic Acids Res*, 36:D637–D640, 2008.

[39] K R Brown and I Jurisca. Unequal evolutionary conservation of human protein interactions in interologous networks. *Genome Biol*, 8:R95, 2007.

[40] M P S Brown, W N Grundy, D Lin, N Cristianini, C W Sugnet, T S Furey, M Ares, and D Haussler. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc Natl Acad Sci U S A*, 97(1):262–267, 2000.

[41] L Buday and J Downward. Roles of cortactin in tumor pathogenesis. *Biochim Biophys Acta*, 1775(2):263–273, 2007.

[42] J M Bueno-de Mesquita, W H van Harten, V P Retel, L J van't Veer, F W A M van Dam, K Karsenberg, K F L Douma, H van Tinteren, J L Peterse, J Wesseling, T S Wu, D Atsma, E J T Rutgers, G Brink, A N Floore, A M Glas, R M H Roumen, F E Bellot, C van Krimpen, and *et al.* Use of 70-gene signature to predict prognosis of patients with node-negative breast cancer:

a prospective community-based feasibility study (RASTER). *Lancet Oncol*, 8:1079–1087, 2007.

[43] W Burke, M Daly, J Garber, J Botkin, M J Kahn, P Lynch, A McTiernan, K Offit, J Perlman, G Petersen, E Thomson, and C Varricchio. Recommendations for follow-up care of individuals with an inherited predisposition to cancer. II. BRCA1 and BRCA2. *J Am Med Assoc*, 277:997–1003, 1997.

[44] P R Burton, D G Clayton, L R Cardon, N Craddock, P Deloukas, A Duncanson, D P Kwiatkowski, M I McCarthy, W H Ouwehand, N J Samani, J A Todd, P Donnelly, and *et al.* Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447:661–678, 2007.

[45] M Buyse, S Loi, L van't Veer, G Viale, M Delorenzi, A M Glas, M Saghatchian d'Assignies, J Bergh, R Lidereau, P Ellis, A Harris, J Bogaerts, P Therasse, A Floore, M Amakrane, F Piette, E Rutgers, C Sotiriou, F Cardoso, and M J Piccart. Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer. *J Natl Cancer Inst*, 98(17):1183–1192, 2006.

[46] I Cadron, T Van Gorp, F Amant, I Vergote, P Moerman, E Waelkens, A Daemen, R Van de Plas, B De Moor, and R Zeillinger. The use of laser microdissection and SELDI-TOF MS in ovarian cancer tissue to identify protein profiles. *Anticancer Res*, 29(4):1039–1046, 2009.

[47] I Cadron, T Van Gorp, D Timmerman, F Amant, E Waelkens, and I Vergote. Application of proteomics in ovarian cancer: which sample should be used? *Gynecol Oncol*, 115:497–503, 2009.

[48] F Cardoso, L van't Veer, E Rutgers, S Loi, S Mook, and M J Piccart-Gebhart. Clinical application of the 70-gene profile: the MINDACT trial. *J Clin Oncol*, 26(5):729–735, 2008.

[49] M P Cary, G D Bader, and C Sander. Pathway information for systems biology. *FEBS Lett*, 579:1815–1820, 2005.

[50] R Caspi, T Altman, J M Dale, K Dreher, C A Fulcher, F Gilham, P Kaipa, A S Karthikeyan, A Kothari, M Krummenacker, M Latendresse, L A Mueller, S Paley, L Popescu, A Pujar, A G Shearer, P Zhang, and P D Karp. The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res*, 38:D473–D479, 2010.

[51] G C Cawley. Leave-one-out cross-validation based model selection criteria for weighted LS-SVMs. In *Proc Int Joint Conf Neural Networks*, pages 1661–1668, 2006.

[52] A Ceol, A Chatr-Aryamontri, L Licata, D Peluso, L Briganti, L Perfetto, L Castagnoli, and G Cesareni. MINT, the Molecular INTeraction database: 2009 update. *Nucleic Acids Res*, 38:D532–D539, 2010.

[53] B Chandrasekaran and A K Jain. Quantization complexity and independent measurements. *IEEE Trans Comput*, 23:102–106, 1974.

[54] 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. *Proc Natl Acad Sci USA*, 102(10):3738–3743, 2005.

[55] H Chen, E Rubin, H Zhang, S Chung, C C Jie, E Garrett, S Biswal, and S Sukumar. Identification of transcriptional targets of HOXA5. *J Biol Chem*, 2809(19):19373–19380, 2005.

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

[57] L Chin and J W Gray. Translating insights from the cancer genome into clinical practice. *Nature*, 452:553–563, 2008.

[58] W C S Cho. Contribution of oncoproteomics to cancer biomarker discovery. *Mol Cancer*, 6:25, 2007.

[59] H-Y Chuang, E Lee, Y-T Liu, D Lee, and T Ideker. Network-based classification of breast cancer metastasis. *Mol Syst Biol*, 3:140, 2007.

[60] C H Chung, S Levy, P Chaurand, and D P Carbone. Genomics and proteomics: emerging technologies in clinical cancer research. *Crit Rev Oncol Hematol*, 61(1):1–25, 2007.

[61] G M Church. The personal genome project. *Mol Syst Biol*, 1:2005.0030, 2005.

[62] F Ciardiello and G Tortora. Epidermal growth factor receptor (EGFR) as a target in cancer therapy: understanding the role of receptor expression and other molecular determinants that could influence the response to anti-EGFR drugs. *Eur J Cancer*, 39:1348–1354, 2003.

[63] D P Clark. Proteomics: the global analysis of proteins. In *Molecular biology: understanding the genetic revolution*, pages 718–719. Elsevier Academic Press, London, 1st edition, 2001.

[64] S Cleator, T J Powles, T Dexter, L Fulford, A Mackay, I E Smith, H Valgeirsson, A Ashworth, and M Dowsett. The effect of the stromal component of breast tumours on prediction of clinical outcome using gene expression microarray analysis. *Breast Cancer Res*, 8:R32, 2006.

[65] J Climent, J L Garcia, J H Mao, J Arsuaga, and J Perez-Losada. Characterization of breast cancer by array comparative genomic hybridization. *Biochem Cell Biol*, 85:497–508, 2007.

[66] G Condous, E Okaro, A Khalid, D Timmerman, C Lu, Y Zhou, S Van Huffel, and T Bourne. The use of a new logistic regression model for predicting the outcome of pregnancies of unknown location. *Hum Reprod*, 19:1900–1910, 2004.

[67] T P Conrads, V A Fusaro, S Ross, D Johann, V Rajapakse, B A Hitt, S M Steinberg, E C Kohn, D A Fishman, G Whiteley, J C Barrett, L A Liotta, E F Petricoin, and T D Veenstra. High-resolution serum proteomic features for ovarian cancer detection. *Endocr Relat Cancer*, 11:163–178, 2004.

[68] International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409:860–921, 2001.

[69] The ENCODE Project Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 447:799–816, 2007.

[70] G M Cooper, T Zerr, J M Kidd, E E Eichler, and D A Nickerson. Systematic assessment of copy number variant detection via genome-wide SNP genotyping. *Nat Genet*, 40:1199–1203, 2008.

[71] N Cristianini and J Shawe-Taylor. *An introduction to Support Vector Machines and other kernel-based learning methods.* Cambridge University Press, Cambridge, 2000.

[72] D S Cross and J K Burmester. Functional characterization of the GDEP promoter and three enhancer elements in retinoblastoma and prostate cell lines. *Med Oncol*, 25(1):40–49, 2008.

[73] N A Cross, S Chandrasekharan, N Jokonya, A Fowles, F C Hamdy, D J Buttle, and C L Eaton. The expression and regulation of ADAMTS-1, -4, -5, -9, and -15, and TIMP-3 by TGFbeta1 in prostate cells: relevance to the accumulation of versican. *Prostate*, 63(3):269–275, 2005.

[74] J A Cruz and D S Wishart. Applications of machine learning in cancer prediction and prognosis. *Cancer Inform*, 2:59–78, 2006.

[75] A Daemen, C Bottomley, O Gevaert, B De Moor, D Timmerman, and T Bourne. Predicting early pregnancy loss with Functional Linear Discriminant Analysis (FLDA). *Ultrasound Obstet Gynecol*, 30(4):469, 2007.

[76] A Daemen, C Bottomley, F Mukri, E Kirk, A T Papageorghiou, D Timmerman, B De Moor, and T Bourne. A new statistical method for longitudinal assessment of early fetal growth - prospective testing of a functional linear discriminant analysis model. *Hum Reprod*, 23(1):i43–i45, 2008.

[77] A Daemen and B De Moor. Development of a kernel function for clinical data. In *Conf Proc IEEE Eng Med Biol Soc*, pages 5913–5917, 2009.

[78] A Daemen, O Gevaert, T De Bie, A Debucquoy, J-P Machiels, B De Moor, and K Haustermans. Integrating microarray and proteomics data to predict the response on cetuximab in patients with rectal cancer. In *Pacific Symposium on Biocomputing*, pages 166–177, 2008.

[79] A Daemen, O Gevaert, and B De Moor. Integration of clinical and microarray data with kernel methods. In *Conf Proc IEEE Eng Med Biol Soc*, pages 5411–5415, 2007.

[80] A Daemen, O Gevaert, K Leunen, E Legius, I Vergote, and B De Moor. Supervised classification of array CGH data with HMM-based feature selection. In *Pacific Symposium on Biocomputing*, pages 468–479, 2009.

[81] A Daemen, O Gevaert, K Leunen, V Vanspauwen, G Michils, E Legius, I Vergote, and B De Moor. Classification of sporadic and BRCA1 ovarian cancer based on a genome-wide study of copy number variations. In I Lovrek, R J Howlett, and L C Jain, editors, *Knowledge-Based Intelligent Information and Engineering Systems*, pages 165–172, 2008.

[82] A Daemen, O Gevaert, K Leunen, V Vanspauwen, G Michils, E Legius, I Vergote, and B De Moor. A genome-wide computational study of copy number variations: an example on ovarian cancer. In A Vellido and P Lisboa, editors, *Investigating human cancer with computational intelligence techniques*, pages 107–118. Future Technology Press, Shoreham-by-sea, UK, 1st edition, 2009.

[83] A Daemen, O Gevaert, F Ojeda, A Debucquoy, J A K Suykens, C Sempoux, J-P Machiels, K Haustermans, and B De Moor. A kernel-based integration of genome-wide data for clinical decision support. *Genome Med*, 1:39, 2009.

[84] A Daemen, O Gevaert, F Ojeda, J A K Suykens, and B De Moor. HI-DID-IT, a HIgh-DImensional Data Integration Toolbox for clinical applications. Submitted.

[85] A Daemen, D Jurkovic, C Van Holsbeke, S Guerriero, A C Testa, A Czekierdowski, R Fruscio, D Paladini, P Neven, A Rossi, T Bourne, B De Moor, and D Timmerman. The effect of cancer prevalence on the use of different cut-off levels and the performance of mathematical models to distinguish benign from malignant adnexal masses. Submitted.

[86] A Daemen, M Signoretto, O Gevaert, J A K Suykens, and B De Moor. Improved microarray-based decision support with graph encoded interactome data. *PLoS ONE*, 5(4):e10225, 2010.

[87] A Daemen, D Timmerman, T Van den Bosch, C Bottomley, E Kirk, C Van Holsbeke, L Valentin, T Bourne, and B De Moor. Improved modeling of clinical data with kernel methods. Submitted.

[88] A Daemen, L Valentin, R Fruscio, C Van Holsbeke, G B Melis, S Guerriero, A Czekierdowski, D Jurkovic, W Ombelet, A Rossi, I Vergote, T Bourne, B De Moor, and D Timmerman. Improving the preoperative classification of adnexal masses as benign or malignant with second stage tests when the basic test is subjective evaluation of ultrasound findings or a mathematical formula for the calculation of risk of malignancy. Submitted.

[89] M Dai, P Wang, A D Boyd, G Kostov, B Athey, E G Jones, W E Bunney, R M Myers, T P Speed, H Akil, S J Watson, and F Meng. Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res*, 33:e175, 2005.

[90] T Damoulos and M A Girolami. Probabilistic multi-class multi-kernel learning: on protein fold recognition and remote homology detection. *Bioinformatics*, 10:1264–1270, 2008.

[91] J Davies, I Wilson, and W Lam. Array CGH technologies and their applications to cancer genomes. *Chromosome Res*, 13:237–248, 2005.

[92] T De Bie. *Semi-supervised learning based on kernel methods and graph cut algorithms.* PhD thesis, Katholieke Universiteit Leuven, Department of Electrical Engineering, 2005.

[93] T De Bie, L C Tranchevent, L M M van Oeffelen, and Y Moreau. Kernel-based data fusion for gene prioritization. *Bioinformatics*, 23:i125–i132, 2007.

[94] A Debucquoy, K Haustermans, A Daemen, S Aydin, L Libbrecht, O Gevaert, B De Moor, S Tejpar, W H McBride, F Penninckx, P Scalliet, C Stroh, S Vlassak, C Sempoux, and J-P Machiels. Molecular response to cetuximab and efficacy of preoperative cetuximab-based chemoradiation in rectal cancer. *J Clin Oncol*, 27:2751–2757, 2009.

[95] J Demšar. Statistical comparisons of classifiers over multiple data sets. *J Mach Learn Res*, 7:1–30, 2006.

[96] A Descazeaud, A de la Taille, Y Allory, H Faucon, L Salomon, T Bismar, R Kim, M D Hofer, D Chopin, C C Abbou, and M A Rubin. Characterization of ZAG protein expression in prostate cancer using a semi-automated microscope system. *Prostate*, 66(10):1037–1043, 2006.

[97] C Desmedt, F Piette, S Loi, Y Wang, F Lallemand, B Haibe-Kains, G Viale, M Delorenzi, Y Zhang, M Saghatchian d'Assignies, J Bergh, R Lidereau, P Ellis, A L Harris, J G M Klijn, J A Foekens, F Cardoso, M J Piccart, M Buyse, and C Sotiriou. Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter independent validation series. *Clin Cancer Res*, 13(11):3207–3214, 2007.

[98] C Desmedt and C Sotiriou. Proliferation: the most prominent predictor of clinical outcome in breast cancer. *Cell Cycle*, 5:2198–2202, 2006.

[99] J K Diss, D J Faulkes, M M Walker, A Patel, C S Foster, V Budhram-Mahadeo, M B Djamgoz, and D S Latchman. Brn-3a neuronal transcription factor functional expression in human prostate cancer. *Prostate Cancer Prostatic Dis*, 9(1):83–91, 2006.

[100] Z Dong, Y Liu, S Lu, A Wang, K Lee, L H Wang, M Revelo, and S Lu. Vav3 oncogene is overexpressed and regulates cell growth and androgen receptor activity in human prostate cancer. *Mol Endocrinol*, 20(10):2315–2325, 2006.

[101] S Dudoit, Y H Yang, M Callow, and T Speed. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statist Sinica*, 12:111–139, 2002.

[102] R P W Duin. The combining classifier: to train or not to train? In *Conf Proc Patt Recogn*, pages 765–770, 2002.

[103] R P W Duin and D M J Tax. Experiments with classifier combining rules. In J Kittler and F Roli, editors, *Multiple Classifier Systems*, pages 16–29. Springer-Verlag, Heidelberg, 2000.

[104] D Dunkler, S Michiels, and M Schemper. Gene expression profiling: does it add predictive accuracy to clinical characteristics in cancer prognosis? *Eur J Cancer*, 43:745–751, 2007.

[105] A Dupuy and R M Simon. Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. *J Natl Cancer Inst*, 99:147–157, 2007.

[106] A Dutt and R Beroukhim. Single nucleotide polymorphism array analysis of cancer. *Curr Opin Oncol*, 19:43–49, 2007.

[107] P Edén, C Ritz, C Rose, M Fernö, and C Peterson. "Good old" clinical markers have similar power in breast cancer prognosis as microarray gene expression profilers. *Eur J Cancer*, 40:1837–1841, 2004.

[108] M Ehrich, M R Nelson, P Stanssens, M Zabeau, T Liloglou, G Xinarianos, C R Cantor, J K Field, and D van den Boom. Quantitative high-throughput analysis of DNA methylation patterns by base-specific cleavage and mass spectrometry. *Proc Natl Acad Sci U S A*, 102(44):15785–15790, 2005.

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

[110] E A Eksioglu, S S Mahmood, M Chang, and V Reddy. GM-CSF promotes differentiation of human dendritic cells and T lymphocytes toward a predominantly type 1 proinflammatory response. *Exp Hematol*, 35:1163–1171, 2007.

[111] R Engers, M Mueller, A Walter, J G Collard, R Willers, and H E Gabbert. Prognostic relevance of Tiam1 protein expression in prostate carcinomas. *Br J Cancer*, 95(8):1081–1086, 2006.

[112] A Esquela-Kerscher and F J Slack. Oncomirs - microRNAs with a role in cancer. *Nat Rev Cancer*, 6:259–269, 2006.

[113] M Esteller. Epigenetic gene silencing in cancer: the DNA hypermethylome. *Hum Mol Genet*, 16(1):R50–R59, 2007.

[114] M Esteller. Epigenetics in cancer. *N Engl J Med*, 358(11):1148–1159, 2008.

[115] M J Fackler, M McVeigh, J Mehrotra, M A Blum, J Lange, A Lapides, E Garrett, P Argani, and S Sukumar. Quantitative multiplex methylation-specific PCR assay for the detection of promoter hypermethylation in multiple genes in breast cancer. *Cancer Res*, 64(13):4442–4452, 2004.

[116] X Fan, L Shi, H Fang, Y Cheng, R Perkins, and W Tong. DNA microarrays are predictive of cancer prognosis: a re-evaluation. *Clin Cancer Res*, 16(2):629–636, 2010.

[117] J Ferlay, P Autier, M Boniol, M Heanue, M Colombet, and P Boyle. Estimates of the cancer incidence and mortality in Europe in 2006. *Ann Oncol*, 18:581–592, 2007.

[118] P Ferroni, R Palmirotta, A Spila, F Martini, V Raparelli, E Fossile, S Mariotti, G Del Monte, O Buonomo, M Roselli, and F Guadagni. Prognostic significance of adiponectin levels in non-metastatic colorectal cancer. *Anticancer Res*, 27(1B):483–489, 2007.

[119] R D Finn, J Mistry, J Tate, P C Coggill, A Heger, J E Pollington, O L Gavin, P Gunasekaran, G Ceric, K Forslund, L Holm, E L Sonnhammer, S R Eddy, and A Bateman. The Pfam protein families database. *Nucleic Acids Res*, 38:D211–D222, 2010.

[120] J A Foekens, D Atkins, Y Zhang, F C Sweep, N Harbeck, A Paradiso, T Cufer, A M Sieuwerts, D Talantov, P N Span, V C Tjan-Heijnen, A F Zito, K Specht, H Hoefler, R Golouh, F Schittulli, M Schmitt, L V Beex, J G Klijn, and Y Wang. Multicenter validation of a gene expression-based prognostic signature in lymph node-negative primary breast cancer. *J Clin Oncol*, 21:171–178, 2006.

[121] S A Forbes, G Bhamra, S Bamford, E Dawson, C Kok, J Clements, A Menzies, J W Teague, P A Futreal, and M R Stratton. The catalogue of somatic mutations in cancer (COSMIC). *Curr Protoc Hum Genet*, 10:10.11, 2008.

[122] F Fouss, A Pirotte, J-M Renders, and M Saerens. Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 19:355–369, 2007.

[123] P J French, J Peeters, S Horsman, E Duijm, I Siccama, M J van den Bent, T M Luider, J M Kros, P van der Spek, and P A Sillevis Smitt. Identification of differentially regulated splice variants and novel exons in glial brain tumors using exon expression arrays. *Cancer Res*, 67(12):5635–5642, 2007.

[124] J Fridlyand, A M Snijders, B Ylstra, H Li, A Olshen, R Segraves, S Dairkee, T Tokuyasu, B M Ljung, A N Jain, J McLennan, J Ziegler, K Chin, S Devries, H Feiler, J W Gray, F Waldman, D Pinkel, and D G Albertson. Breast tumor copy number aberration phenotypes and genomic instability. *BMC Cancer*, 6:96, 2006.

[125] J H Friedman. Regularized discriminant analysis. *J Am Stat Ass*, 84:165–175, 1989.

[126] R C Friedman, K Farh, C B Burge, and D P Bartel. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res*, 19:92–105, 2009.

[127] S Fröhling and H Döhner. Chromosomal abnormalities in cancer. *N Engl J Med*, 359(7):722–734, 2008.

[128] T S Furey, N Cristianini, N Duffy, D W Bednarski, M Schummer, and D Haussler. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, 16(2):906–914, 2000.

[129] B Furusato, CL Gao, L Ravindranath, Y Chen, J Cullen, D G McLeod, A Dobi, S Srivastava, G Petrovics, and I A Sesterhenn. Mapping of TMPRSS2-ERG fusions in the context of multi-focal prostate cancer. *Mod Pathol*, 21(2):67–75, 2008.

[130] P A Futreal, L Coin, M Marshall, T Down, T Hubbard, R Wooster, N Rahman, and M R Stratton. A census of human cancer genes. *Nature rev*, 4:117–183, 2004.

[131] M E Futschik, M Sullivan, A Reeve, and N Kasabov. Prediction of clinical behaviour and treatment for cancers. *Appl Bioinf*, 2:S53–S58, 2003.

[132] W Gajewski and R D Legare. Ovarian cancer. *Surg Oncol Clin N Am*, 7:317–333, 1998.

[133] M Y Galperin and G R Cochrane. Nucleic Acids Research annual database issue and the NAR online molecular biology database collection in 2009. *Nucleic Acids Res*, 37:D1–D4, 2009.

[134] C Garnis, W W Lockwood, E Vucic, Y Ge, L Girard, J D Minna, A F Gazdar, S Lam, C MacAulay, and W L Lam. High resolution analysis of non-small cell lung cancer cell lines by whole genome tiling path array CGH. *Int J Cancer*, 118:1556–1564, 2006.

[135] L A Garraway, H R Widlund, M A Rubin, G Getz, A J Berger, S Ramaswamy, R Beroukhim, D A Milner, S R Granter, J Du, C Lee, S N Wagner, C Li, T R Golub, D L Rimm, M L Meyerson, D E Fisher, and W R Sellers. Integrative genomic analyses identify MITF as a lineage survival oncogene amplified in malignant melanoma. *Nature*, 436:117–122, 2005.

[136] M G Genton. Classes of kernels for machine learning: a statistics perspective. *J Mach Learn Res*, 2:299–312, 2001.

[137] O Gevaert. *A Bayesian network integration framework for modeling biomedical data.* PhD thesis, Katholieke Universiteit Leuven, Department of Electrical Engineering, 2008.

[138] O Gevaert, A Daemen, B De Moor, and L Libbrecht. A taxonomy of epithelial human cancer and their metastasis. *BMC Med Genomics*, 2:69, 2009.

[139] O Gevaert, F De Smet, E Kirk, B Van Calster, T Bourne, S Van Huffel, Y Moreau, D Timmerman, B De Moor, and G Condous. Predicting the outcome of pregnancies of unknown location: Bayesian networks with expert prior information compared to logistic regression. *Hum Reprod*, 21:1824–1831, 2006.

[140] F C Geyer and J S Reis-Filho. Microarray-based gene expression profiling as a clinical tool for breast cancer management: are we there yet? *Int J Surg Path*, 17(4):285–302, 2009.

[141] D F Gleason. Classification of prostatic carcinomas. *Cancer Chemother Rep*, 50(3):125–128, 1966.

[142] C Goble and R Stevens. State of the nation in data integration for bioinformatics. *J Biom Inf*, 41:687–693, 2008.

[143] T R Golub, D K Slonim, P Tamayo, C Huard, M Gaasenbeek, J P Mesirov, H Coller, M L Loh, J R Downing, M A Caligiuri, C D Bloomfield, and E S Lander. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537, 1999.

[144] L A M Gravendeel, M C M Kouwenhoven, O Gevaert, J J de Rooi, A P Stubbs, J E Duijm, A Daemen, F E Bleeker, L B C Bralten, N K Kloosterhof, B De Moor, P H Eilers, P J van der Spek, J M Kros, P A E Sillevis Smitt, M J van den Bent, and P J French. Intrinsic gene expression profiles of gliomas are a better predictor of survival than histology. *Cancer Res*, 69(23):9065–9072, 2009.

[145] J W Gray. Evidence emerges for early metastasis and parallel evolution of primary and metastatic tumors. *Cancer Cell*, 4:4–6, 2003.

[146] W J Griffiths and Y Wang. Mass spectrometry: from proteomics to metabolomics and lipidomics. *Chem Soc Rev*, 38(7):1882–1896, 2009.

[147] S Griffiths-Jones, H K Saini, S van Dongen, and A J Enright. miRBase: tools for microRNA genomics. *Nucleic Acids Res*, 36:D154–D158, 2008.

[148] J Gröne, B Weber, E Staub, M Heinze, I Klaman, C Pilarsky, K Hermann, E Castanos-Velez, S Röpcke, B Mann, A Rosenthal, and H J Buhr. Differential expression of genes encoding tight junction proteins in colorectal cancer: frequent dysregulation of claudin-1, -8 and -12. *Int J Colorectal Dis*, 22:651–659, 2007.

[149] S K Gruvberger-Saal, H E Cunliffe, K M Carr, and I A Hedenfalk. Microarrays in breast cancer research and clinical practice - the future lies ahead. *Endocr Relat Cancer*, 13:1017–1031, 2006.

[150] O Habayeb, A Daemen, D Timmerman, B De Moor, G A Hackett, T Bourne, and C C Lees. The relationship between first trimester fetal growth, pregnancy associated plasma protein A levels and birthweight. Submitted.

[151] L A Habel, S Shak, M K Jacobs, A Capra, C Alexander, M Pho, J Baker, M Walker, D Watson, J Hackett, N T Blick, D Greenberg, L Fehrenbacher, B Langholz, and C P Quesenberry. A population-based study of tumor gene expression and risk of breast cancer death among lymph node-negative patients. *Breast Cancer Res*, 8:R25, 2006.

[152] S Hanash, S Pitteri, and V Faca. Mining the plasma proteome for cancer biomarkers. *Nature*, 452:571–579, 2008.

[153] D Hanisch, A Zien, R Zimmer, and T Lengauer. Co-clustering of biological networks and gene expression data. *Bioinformatics*, 18:S145–S154, 2002.

[154] J A Hanley and B J McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143:29–36, 1982.

[155] J A Hanley and B J McNeil. A method of comparing the areas under receiver operating characteristics curves derived from the same cases. *Radiology*, 148:839–843, 1983.

[156] P C Hansen. Numerical tools for analysis and solution of fredholm integral equations of the first kind. *Inverse Probl*, 8:849–872, 1992.

[157] S Haupt, M Berger, Z Goldberg, and Y Haupt. Apoptosis - the p53 network. *J Cell Sci*, 116:4077–4085, 2003.

[158] Z Herceg. Epigenetics and cancer: towards an evaluation of the impact of environment and dietary factors. *Mutagenesis*, 22(2):91–103, 2007.

[159] K R Hess, K Anderson, W F Symmans, V Valero, N Ibrahim, J A Mejia, D Booser, R L Theriault, A U Buzdar, P J Dempsey, R Rouzier, N Sneige, J S Ross, T Vidaurre, H L Gómez, G N Hortobagyi, and L Pusztai. Pharmacogenomic predictor of sensitivity to preoperative chemotherapy with paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide in breast cancer. *J Clin Oncol*, 24:4236–4244, 2006.

[160] L G Horvath, J E Lelliott, J G Kench, C S Lee, E D Williams, D N Saunders, J J Grvgiel, R L Sutherland, and S M Henshall. Secreted frizzled-related protein 4 inhibits proliferation and metastatic potential in prostate cancer. *Prostate*, 67(10):1081–1090, 2007.

[161] C W Hsu, C C Chang, and C J Lin. A practical guide to support vector classification. Posted at http://www.csie.ntu.edu.tw/ cjlin/papers/guide/guide.pdf.

[162] S D Hsu, C H Chu, A P Tsou, S J Chen, H C Chen, P W Hsu, Y H Wong, Y H Chen, G H Chen, and H D Huang. miRNAMap 2.0: genomic maps of microRNAs in metazoan genomes. *Nucleic Acids Res*, 36:D165–D169, 2008.

[163] E Huang, S H Cheng, H Dressman, J Pittman, M H Tsou, C F Horng, A Bild, E S Iversen, M Liao, C M Chen, M West, J R Nevins, and A T Huang. Gene expression predictors of breast cancer outcomes. *Lancet*, 361:1590–1596, 2003.

[164] T J P Hubbard, B L Aken, S Ayling, B Ballester, K Beal, E Bragin, S Brent, Y Chen, P Clapham, L Clarke, G Coates, S Fairley, S Fitzgerald, J Fernandez-Banet, L Gordon, S Graf, S Haider, M Hammond, R Holland, K Howe, A Jenkinson, N Johnson, A Kahari, D Keefe, S Keenan, and *et al.* Ensembl 2009. *Nucleic Acids Res*, 37:D690–D697, 2009.

[165] T Hudolin, A Juretic, G C Spagnoli, J Pasini, D Bandic, M Heberer, M Kosicek, and M Cacic. Immunohistochemical expression of tumor antigens MAGE-A1, MAGE-A3/4, and NY-ESO-1 in cancerous and benign prostatic tissue. *Prostate*, 66(11):13–18, 2006.

[166] D L Hunt, R B Haynes, S E Hanna, and K Smith. Effects of computer-based clinical decision support systems on physician performance and patient outcomes. *J Am Med Assoc*, 280:1339–1346, 1998.

[167] T Ideker, T Galitski, and L Hood. A new approach to decoding life: systems biology. *Annu Rev Genomics Hum Genet*, 2:343–372, 2001.

[168] J P A Ioannidis. Microarrays and molecular research: noise discovery? *Lancet*, 365:454–455, 2005.

[169] R A Irizarry, B Hobbs, F Collin, Y D Beazer-Barclay, K J Antonellis, U Scherf, and T Speed. Exploration, normalization and summaries of high density oligonucleotide array probe level data. *Biostat*, 4:249–264, 2003.

[170] K Ishii, S Usui, Y Sugimura, S Yoshida, T Hioki, M Tatematsu, H Yamamoto, and K Hirano. Aminopeptidase N regulated by zinc in human prostate participates in tumor cell invasion. *Int J Cancer*, 92:49–54, 2001.

[171] A V Ivshina, J George, O Senko, B Mow, T C Putti, J Smeds, T Lindahl, Y Pawitan, P Hall, H Nordgren, J E Wong, E T Liu, J Bergh, V A Kuznetsov, and L D Miller. Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer. *Cancer Res*, 66:10292–10301, 2006.

[172] A K Jain, R P W Duin, and J Mao. Statistical pattern recognition: a review. *IEEE Trans Pattern Anal Mach Intell*, 22:4–37, 2000.

[173] G James and T Hastie. Functional linear discriminant analysis for irregularly sampled curves. *J R Stat Soc Series B Stat Methodol*, 63:533–550, 2001.

[174] L J Jensen, M Kuhn, M Stark, S Chaffron, C Creevey, J Muller, T Doerks, P Julien, A Roth, M Simonovic, P Bjork, and C von Mering. STRING 8 - a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res*, 37:D412–D416, 2009.

[175] T-K Jenssen, A Laegreid, J Komorowski, and E Hovig. A literature network of human genes for high-throughput analysis of gene expression. *Nat Genet*, 28:21–28, 2001.

[176] J Jin. Impossibility of successful classification when useful features are rare and weak. *Proc Natl Acad Sci U S A*, 106(22):8859–8864, 2009.

[177] A Kahvejian, J Quackenbush, and J F Thompson. What would you do if you could sequence everything? *Nat Biotechnol*, 26:1125–1133, 2008.

[178] M Kanehisa, S Goto, M Furumichi, M Tanaba, and M Hirakawa. KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res*, 38:D355–D360, 2010.

[179] T Kato, K Tsuda, and K Asai. Selective integration of multiple biological data for supervised network inference. *Bioinformatics*, 21(10):2488–2495, 2005.

[180] T S Keshava Prasad, R Goel, K Kandasamy, S Keerthikumar, S Kumar, S Mathivanan, D Telikicherla, R Raju, B Shafreen, A Venugopal, L Balakrishnan, A Marimuthu, S Banerjee, D S Somanathan, A Sebastian, S Rani, S Ray, C J Harrys Kishore, S Kanth, M Ahmed, and *et al.* Human Protein Reference Database - 2009 update. *Nucleic Acids Res*, 37:D767–D772, 2009.

[181] T D Kim, K S Song, G Li, H Choi, H D Park, K Lim, B D Hwang, and W H Yoon. Activity and expression of urokinase-type plasminogen activator and matrix metalloproteinases in human colorectal cancer. *BMC Cancer*, 6:211, 2006.

[182] E Kirk, A Daemen, A T Papageourghiou, C Bottomley, G Condous, B De Moor, D Timmerman, and T Bourne. Why are some ectopic pregnancies characterized as pregnancies of unknown location at the initial transvaginal ultrasound examination. *Acta Obstet Gynecol Scand*, 87(11):1150–1154, 2008.

[183] J Kittler. Combining classifiers: a theoretical framework. *Pattern Anal Appl*, 1:18–27, 1998.

[184] C Klijn, H Holstege, J de Ridder, X Liu, M Reinders, J Jonkers, and L Wessels. Identification of cancer genes using a statistical framework for multiexperiment analysis of nondiscretized array CGH data. *Nucleic Acids Res*, 36(2):e13, 2008.

[185] M Kloor, S Michel, B Buckowitz, J Ruschoff, R Buttner, E Holinski-Feder, W Dippold, R Wagner, M Tariverdian, A Benner, Y Schwitalle, B Kuchenbuch, and M von Knebel Doeberitz. Beta2-microglobulin mutations in microsatellite unstable colorectal tumors. *Int J Cancer*, 121(2):454–458, 2007.

[186] W Kolch, H Mischak, and A R Pitt. The molecular make-up of a tumour: proteomics in cancer research. *Clin Sci*, 108:369–383, 2005.

[187] R I Kondor and J Lafferty. Diffusion kernels on graphs and other discrete structures. In *Proc Int Conf Mach Learn*, pages 315–322, 2002.

[188] I Kononenko. Machine learning for medical diagnosis: history, state of the art and perspective. *Art Intell Med*, 23:89–109, 2001.

[189] S Koscielny. Critical review of microarray-based prognostic tests and trials in breast cancer. *Curr Opin Obstet Gynecol*, 20:47–50, 2008.

[190] A Kretschmer, K Moepert, S Dames, M Sternberger, J Kaufmann, and A Klippel. Differential regulation of TGF-beta signaling through Smad2, Smad3 and Smad4. *Oncogene*, 22(43):6748–6763, 2003.

[191] M Kuhn, D Szklarczyk, A Franceschini, M Campillos, C von Mering, L J Jensen, A Beyer, and P Bork. STITCH 2: an interaction network database for small molecules and proteins. *Nucleic Acids Res*, 38:D552–D556, 2010.

[192] L Kummola, J M Hämäläinen, J M Kivelä, A J Kivelä, J Saarnio, T Karttunen, and S Parkkila. Expression of a novel carbonic anhydrase, CA XIII, in normal and neoplastic colorectal mucosa. *BMC Cancer*, 5:41, 2005.

[193] T Kwan, D Benovoy, C Dias, S Gurd, C Provencher, P Beaulieu, T J Hudson, R Sladek, and J Majewski. Genome-wide analysis of transcript isoform variation in humans. *Nat Genet*, 40:225–231, 2008.

[194] 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. *Bioinformatics*, 7:235–244, 2006.

[195] W R Lai, M D Johnson, R Kucherlapati, and P J Park. Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. *Bioinformatics*, 21(19):3763–3770, 2005.

[196] G R G Lanckriet, N Cristianini, P Bartlett, L El Ghaoui, and M I Jordan. Learning the kernel matrix with semidefinite programming. *J Mach Learn Res*, 5:27–72, 2004.

[197] G R G Lanckriet, T De Bie, N Cristianini, M I Jordan, and W S Noble. A statistical framework for genomic data fusion. *Bioinformatics*, 20(16):2626–2635, 2004.

[198] J Lapointe, C Li, C P Giacomini, K Salari, S Huang, P Wang, M Ferrari, T Hernandez-Boussard, J D Brooks, and J R Pollack. Genomic profiling reveals alternative genetic pathways of prostate tumorigenesis. *Cancer Res*, 67(18):8504–8510, 2007.

[199] J Lapointe, C Li, J P Higgins, M van de Rijn, E Bair, K Montgomery, M Ferrari, L Egevad, W Rayford, U Bergerheim, P Ekman, A M DeMarzo, R Tibshirani, D Botstein, P O Brown, J D Brooks, and J R Pollack. Gene expression profiling identifies clinically relevant subtypes of prostate cancer. *Proc Natl Acad Sci U S A*, 101(3):811–816, 2004.

[200] P Larrañaga, B Calvo, R Santana, C Bielza, J Galdiano, I Inza, J A Lozano, R Armañanzas, G Santafé, A Pérez, and V Robles. Machine learning in bioinformatics. *Brief Bioinform*, 7(1):86–112, 2006.

[201] H Ledford. The death of microarrays? *Nature*, 455:847, 2008.

[202] E Lee, H Y Chuang, J W Kim, T Ideker, and D Lee. Inferring pathway activity toward precise disease classification. *PLoS Comput Biol*, 4:e1000217, 2008.

[203] W Lei, J J Rushton, L M Davis, F Liu, and S A Ness. Positive and negative determinants of target gene specificity in Myb transcription factors. *J Biol Chem*, 279(28):29519–29527, 2004.

[204] K Leunen, O Gevaert, A Daemen, V Vanspauwen, G Michils, B De Moor, P Moerman, I Vergote, and E Legius. Recurrent copy number alterations in BRCA1-mutated ovarian tumors alter biological pathways. *Hum Mutat*, 30(12):1693–1702, 2009.

[205] C Li and H Li. Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics*, 24:1175–1182, 2008.

[206] W Li and Y Yang. How many genes are needed for a discriminant microarray data analysis. In S M Lin and K F Johnson, editors, *Methods of Microarray Data Analysis*, pages 137–150. Kluwer Academic, 2002.

[207] R P Lifton. Individual genomes on the horizon. *New Engl J Med*, page NEJMe1001090, 2010.

[208] J Liu, S Ranka, and T Kahveci. Classification and feature selection algorithms for multi-class CGH data. *Bioinformatics*, 24(13):i86–i95, 2008.

[209] J Lohi. Laminin-5 in the progression of carcinomas. *Int J Cancer*, 94:763–767, 2001.

[210] J Louhimo, M Carpelan-Holmström, H Alfthan, U H Stenman, H J Jarvinen, and C Haglund. Serum HCG beta, CA 72-4 and CEA are independent prognostic factors in colorectal cancer. *Int J Cancer*, 101(6):545–548, 2002.

[211] C Lu, T Van Gestel, J A K Suykens, S Van Huffel, I Vergote, and D Timmerman. Preoperative prediction of malignancy of ovarium tumor using least squares support vector machines. *Artif Intell Med*, 28:281–306, 2003.

[212] J Lu, G Getz, E A Miska, E Alvarez-Saavedra, J Lamb, D Peck, A Sweet-Cordero, B L Ebert, R H Mak, A A Ferrando, J R Downing, T Jacks, H R Horvitz, and T R Golub. MicroRNA expression profiles classify human cancers. *Nature*, 435:834–838, 2005.

[213] X Lu and X Zhang. The effect of GeneChip gene definitions on the microarray study of cancers. *BioEssays*, 28:739–746, 2006.

[214] J Lupski. Genomic disorders ten years on. *Genome Med*, 1:42, 2009.

[215] H Ma, A Sorokin, A Mazein, A Selkov, E Selkov, O Demin, and I Goryanin. The Edinburgh human metabolic network reconstruction and its functional analysis. *Mol Syst Biol*, 3:135, 2007.

[216] G MacBeatch and S L Schreiber. Printing proteins as microarrays for high-throughput function determination. *Science*, 289:1760–1763, 2000.

[217] J P Machiels, S Aydin, M A Bonny, F Hammouch, and C Sempoux. What is the best to predict disease-free survival after preoperative radiochemotherapy for rectal cancer patients: tumor regression grading, nodal status or circumferential resection margin invasion? *J Clin Oncol*, 24:1319–1321, 2006.

[218] J-P Machiels, C Sempoux, P Scalliet, J C Coche, Y Humblet, E Van Cutsem, J Kerger, J L Canon, M Peeters, S Aydin, S Laurent, A Kartheuser, B Coster, S Roels, J F Daisne, B Honhon, L Duck, C Kirkove, M A Bonny, and K Haustermans. Phase I/II study of preoperative cetuximab, capecitabine, and external beam radiotherapy in patients with rectal cancer. *Ann Oncol*, 18:738–744, 2007.

[219] K Maeda, S M Kang, T Sawada, Y Nishiguchi, M Yashiro, Y Ogawa, M Ohira, T Ishikawa, Y S Hirakawa, and K Chung. Expression of intercellular adhesion molecule-1 and prognosis in colorectal cancer. *Oncol Rep*, 9(3):511–514, 2002.

[220] C Maihöfner, M P Charalambous, U Bhambra, T Lightfoot, G Geisslinger, and N J Gooderham. The Colorectal Cancer Group (2003) Expression of cyclooxygenase-2 parallels expression of interleukin-1beta, interleukin-6 and NF-kappaB in human colorectal cancer. *Carcinogenesis*, 24(4):665–671, 2003.

[221] S K Majumder, N Ghosh, and P K Gupta. Relevance vector machine for optical diagnosis of cancer. *Lasers Surg Med*, 36:323–333, 2005.

[222] K M Mani, C Lefebvre, K Wang, W K Lim, K Basso, R Dalla-Favera, and A Califano. A systems biology approach to prediction of oncogenes and molecular perturbation targets in B-cell lymphomas. *Mol Syst Biol*, 4:169, 2008.

[223] J P Mathew, B S Taylor, G D Bader, S Pyarajan, M Antoniotti, A M Chinnaiyan, C Sander, S J Burakoff, and B Mishra. From bytes to bedside: data integration and computational biology for translational cancer research. *PLoS Comput Biol*, 3(2):e12, 2007.

[224] L Matthews, G Gopinath, M Gillespie, M Caudy, D Croft, B de Bono, P Garapati, J Hemish, H Hermjakob, B Jassal, A Kanapin, S Lewis, S Mahajan, B May, E Schmidt, I Vastrik, G Wu, E Birney, L Stein, and P D'Eustachio. Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res*, 37:D619–D622, 2009.

[225] M Mesiti, E Jiménez-Ruiz, I Sanz, R Berlanga-Llavori, P Perlasca, G Valentini, and D Manset. XML-based approaches for the integration of heterogeneous bio-molecular data. *BMC Bioinformatics*, 10(Suppl 12):S7, 2009.

[226] C E Metz. Basic principles of ROC analysis. *Semin Nucl Med*, 8:283–298, 1978.

[227] S Michiels, S Koscielny, and C Hill. Prediction of cancer outcome with microarrays: a multiple random validation strategy. *Lancet*, 365:488–492, 2005.

[228] P Michl, A R Ramjaun, O E Pardo, P H Warne, M Wagner, R Poulsom, C D'Arrigo, K Ryder, A Menke, T Gress, and J Downward. CUTL1 is a target of TGF(beta) signaling that enhances cancer cell motility and invasiveness. *Cancer Cell*, 7(6):521–532, 2005.

[229] L D Miller, J Smeds, J George, V B Vega, L Vergara, A Ploner, Y Pawitan, P Hall, S Klaar, E T Liu, and J Bergh. An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *Proc Natl Acad Sci U S A*, 102:13550–13555, 2005.

[230] O K Mirzoeva, D Das, L M Heiser, S Bhattacharya, D Siwak, R Gendelman, N Bayani, N J Wang, R M Neve, Y Guan, Z Hu, Z Knight, H S Feiler, P Gascard, B Parvin, P T Spellman, K M Shokat, A J Wyrobek, M J Bissell, F McCormick, W L Kuo, G B Mills, J W Gray, and W M Korn. Basal subtype and MAPK/ERK Kinase (MEK)-phosphoinositide 3-kinase feedback signaling determine susceptibility of breast cancer cells to MEK inhibition. *Cancer Res*, 69:565–572, 2009.

[231] K Miyanaga, Y Kato, T Nakamura, M Matsumura, H Amaya, T Horiuchi, Y Chiba, and K Tanaka. Expression and role of thrombospondin-1 in colorectal cancer. *Anticancer Res*, 22(6C):3941–3948, 2002.

[232] V K Mootha, P Lepage, K Miller, J Bunkenborg, M Reich, M Hjerrild, T Delmonte, A Villeneuve, R Sladek, F Xu, G A Mitchell, C Morin, M Mann, T J Hudson, B Robinson, J D Rioux, and E S Lander. Identification of a gene causing human cytochrome c oxidase deficiency by integrative genomics. *Proc Natl Acad Sci U S A*, 100:605–610, 2003.

[233] M Muc-Wierzgon, E Nowakowska-Zajdel, T Kokot, A Kozowicz, B Zubelewicz, K Klakla, U Mazurek, K Cholewa, T Wilczok, J Wierzgon, and K Sosada. Genetic disregulation of gene coding tumor necrosis factor alpha receptors (TNFalpha Rs) in colorectal cancer cells. *J Exp Clin Cancer Res*, 23(4):651–660, 2004.

[234] R K Nam, L Sugar, W Yang, S Srivastava, L H Klotz, L Y Yang, A Stanimirovic, E Encioiu, M Neill, D A Loblaw, J Trachtenberg, S A Narod, and A Seth. Expression of the TMPRSS2:ERG fusion gene predicts cancer recurrence after surgery for localised prostate cancer. *Br J Cancer*, 97:1690–1695, 2007.

[235] J R Nevins, E S Huang, H Dressman, J Pittman, A T Huang, and M West. Towards integrated clinico-genomic models for personalized medicine: combining gene expression signatures and clinical factors in breast cancer outcomes prediction. *Hum Mol Genet*, 12:153–157, 2003.

[236] J Nielsen and S Oliver. The next wave in metabolome analysis. *Trends Biotechnol*, 23(11):544–546, 2005.

[237] B V North, D Curtis, and P C Sham. A note on the calculation of empirical p values from monte carlo procedures. *Am J Hum Genet*, 71:439–441, 2002.

[238] R L Nussbaum, R R McInnes, and H F Willard. *Thompson & Thompson Genetics in Medicine, 7th Edition.* Saunders Elsevier, Philadelphia, 2007.

[239] C Oakman, S Bessi, E Zafarana, F Galardi, L Biganzoli, and A Di Leo. New diagnostics and biological predictors of outcome in early breast cancer. *Breast Cancer Res*, 11:205, 2009.

[240] S Paik. Development and clinical utility of a 21-gene recurrence score prognostic assay in patients with early breast cancer treated with tamoxifen. *Oncologist*, 12:631–635, 2007.

[241] S Paik, S Shak, G Tang, C Kim, J Baker, M Cronin, F L Baehner, M G Walker, D Watson, T Park, W Hiller, E R Fisher, D L Wickerham, J Bryant, and N Wolmark. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N Engl J Med*, 351:2817–2826, 2004.

[242] S Paik, G Tang, S Shak, C Kim, J Baker, W Kim, M Cronin, F L Baehner, D Watson, J Bryant, J P Costantino, C E Geyer, D L Wickerham, and N Wolmark. Gene expression and benefit of chemotherapy in women with node-negative, estrogen receptor-positive breast cancer. *J Clin Oncol*, 24:3726–3734, 2006.

[243] W Y Park, B Miranda, D Lebeche, G Hashimoto, and W V Cardoso. FGF-10 is a chemotactic factor for distal epithelial buds during lung development. *Dev Biol*, 201(2):125–134, 1998.

[244] P Pavlidis, J Weston, J Cai, and W N Grundy. Learning gene functional classifications from multiple data types. *J Comput Biol*, 9(2):401–411, 2002.

[245] K Pelckmans. *Primal-Dual Kernel Machines.* PhD thesis, Katholieke Universiteit Leuven, Department of Electrical Engineering, 2005.

[246] K Pelckmans, J A K Suykens, T Van Gestel, J De Brabanter, L Lukas, B Hamers, B De Moor, and J Vandewalle. LS-SVMlab: a Matlab/C toolbox for Least Squares Support Vector Machines. Posted at http://www.esat.kuleuven.be/sista/lssvmlab/.

[247] E F Petricoin, A M Ardekani, B A Hitt, P J Levine, V A Fusaro, S M Steinberg, G B Mills, C Simone, D A Fishman, E C Kohn, and L A Liotta. Use of proteomic patterns in serum to identify ovarian cancer. *Lancet*, 359:572–577, 2002.

[248] A Pexsters, A Daemen, C Bottomley, D Van Schoubroeck, L De Catte, B De Moor, T D'Hooghe, C Lees, D Timmerman, and T Bourne. A new crown-rump length size curve based on over 3500 pregnancies. Ultrasound Obstet Gynecol, In press.

[249] A Pexsters, A Daemen, J P Frijns, G Betsas, C Bottomley, D Van Schoubroeck, B De Moor, T D'Hooghe, C Lees, D Timmerman, and T Bourne. Crown rump length in genetically abnormal pregnancies compared to a reference CRL size curve of normal pregnancies. Submitted.

[250] D Pinkel and D G Albertson. Array comparative genomic hybridization and its applications in cancer. *Nat Genet*, 37:S11–S17, 2005.

[251] J Pittman, E Huang, H Dressman, C F Horng, S H Cheng, M H Tsou, C M Chen, A Bild, E S Iversen, A T Huang, J R Nevins, and M West. Integrated modeling of clinical and gene expression information for personalized prediction of disease outcomes. *Proc Natl Acad Sci U S A*, 101:8431–8436, 2004.

[252] C Plass. Cancer epigenomics. *Hum Mol Genet*, 11(20):2479–2488, 2002.

[253] N Pochet. *Microarray data analysis using support vector machines and kernel methods.* PhD thesis, Katholieke Universiteit Leuven, Department of Electrical Engineering, 2006.

[254] N Pochet, F De Smet, J A K Suykens, and B De Moor. Systematic benchmarking of microarray data classification: assessing the role of nonlinearity and dimensionality reduction. *Bioinformatics*, 20:3185–3195, 2004.

[255] N Pochet, F Janssens, F De Smet, K Marchal, J A K Suykens, and B De Moor. M@CBETH: a microarray classification benchmarking tool. *Bioinformatics*, 21(14):3185–3186, 2005.

[256] R Polikar. Ensemble based systems in decision making. *IEEE Circuits and Systems Magazine*, 6(3):21–45, 2006.

[257] C Prieto and J De Las Rivas. APID: Agile Protein Interaction DataAnalyzer. *Nucleic Acids Res*, 34:W298–W302, 2006.

[258] The Human Genome Project. http://www.ornl.gov/sci/techresources/human_genome/home.shtml.

[259] J Quackenbush. Microarray analysis and tumor classification. *New Engl J Med*, 354(23):2463–2472, 2006.

[260] P Quirke, P Durdey, M F Dixon, and N F Williams. Local recurrence of rectal adenocarcinoma due to inadequate surgical resection: histopathological study of lateral tumour spread and surgical excision. *Lancet*, 2:996–999, 1986.

[261] B Raghavachari, A Tasneem, T M Przytycka, and R Jothi. DOMINE: a database of protein domain interactions. *Nucleic Acids Res*, 36:D656–D661, 2008.

[262] D Raje, H Mukhtar, A Oshowo, and C I Clark. What proportion of patients referred to secondary care with iron deficiency anemia have colon cancer? *Dis Colon Rectum*, 50:1–4, 2007.

[263] F Rapaport, E Barillot, and J-P Vert. Classification of arrayCGH data using fused SVM. *Bioinformatics*, 24:i375–i382, 2008.

[264] F Rapaport, A Zinovyev, M Dutreix, E Barillot, and J-P Vert. Classification of microarray data using gene networks. *BMC Bioinformatics*, 8:35, 2007.

[265] S Raudys and R Duin. Expected classification error of the Fisher linear classifier with pseudo-inverse covariance matrix. *Pattern Recognit Lett*, 19:385–392, 1998.

[266] R Redon, S Ishikawa, K R Fitch, L Feuk, G H Perry, T D Andrews, H Fiegler, M H Shapero, A R Carson, W Chen, E K Cho, S Dallaire, J L Freeman, J R González, M Gratacòs, J Huang, D Kalaitzopoulos, D Komura, J R MacDonald, C R Marshall, R Mei, L Montgomery, K Nishimura, and *et al.* Global variation in copy number in the human genome. *Nature*, 444:444–454, 2006.

[267] M M Reinholz, M W An, S A Johnsen, M Subramanian, V J Suman, J N Ingle, P C Roche, and T C Spelsberg. Differential gene expression of TGF beta inducible early gene (TIEG), Smad7, Smad2 and Bard1 in normal and malignant breast tissue. *Breast Cancer Res Treat*, 86(1):75–88, 2004.

[268] I Rish. An empirical study of the naïve Bayes classifier. In *Proc Int Joint Conf Artificial Intelligence*, pages 41–46, 2001.

[269] L Roberts, R J Davenport, E Pennisi, and E Marshall. A history of the human genome project. *Science*, 291:1195, 2001.

[270] H P Robinson. Sonar measurement of fetal crown-rump length as means of assessing maturity in first trimester of pregnancy. *Br Med J*, 4(5883):28–31, 1973.

[271] U Roessner and J Bowne. What is metabolomics all about? *Biotechniques*, 46(5):363–365, 2009.

[272] P Romero, J Wagg, M L Green, D Kaiser, M Krummenacker, and P D Karp. Computational prediction of human metabolic pathways from the complete human genome. *Genome Biol*, 6:R2, 2004.

[273] A Rosenwald, G Wright, W C Chan, J M Connors, E Campo, R I Fisher, R D Gascoyne, H K Muller-Hermelink, E.B Smeland, J M Giltnane, E M Hurt, H Zhao, L Averett, L Yang, W H Wilson, E S Jaffe, R Simon, R D Klausner, J Powell, P L Duffey, D L Longo, T C Greiner, and *et al.* The use of molecular profiling to predict survival after chemotherapy for diffuse large-b-cell lymphoma. *New Engl J Med*, 346:1937–1947, 2002.

[274] J S Ross, C Hatzis, W F Symmans, L Pusztai, and G N Hortobágyi. Commercialized multigene predictors of clinical outcome for breast cancer. *Oncologist*, 13:477–493, 2008.

[275] C Rubie, V O Frick, S Pfeil, M Wagner, O Kollmar, B Kopp, S Graber, B M Rau, and M K Schilling. Correlation of IL-8 with induction, progression and metastatic potential of colorectal cancer. *World J Gastroenterol*, 13(37):4996–5002, 2007.

[276] A Rzhetsky, I Iossifov, T Koike, M Krauthammer, P Kra, M Morris, H Yu, P A Duboué, W Weng, W J Wilbur, V Hatzivassiloglou, and C Friedman. GeneWays: a system for extracting, analyzing, visualizing, and integrating molecular pathway data. *J Biomed Inform*, 37:43–53, 2004.

[277] Y Saeys, I Inza, and P Larrañaga. A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19):2507–2517, 2007.

[278] A Sahni, P J Simpson-Haidaris, S K Sahni, G G Vaday, and C W Francis. Fibrinogen synthesized by cancer cells augments the proliferative effect of fibroblast growth factor-2 (FGF-2). *J Thromb Haemost*, 6(1):176–183, 2008.

[279] L Salwinski, C S Miller, A J Smith, F K Pettit, J U Bowie, and D Eisenberg. The database of interacting proteins: 2004 update. *Nucleic Acids Res*, 32:D449–D451, 2004.

[280] S Santagata, F Demichelis, A Riva, S Varambally, M D Hofer, J L Kutok, R Kim, J Tang, J E Montie, A M Chinnaiyan, M A Rubin, and J C Aster. JAGGED1 expression is associated with prostate cancer metastasis and recurrence. *Cancer Res*, 64(19):6854–6857, 2004.

[281] I Sarris, C Bottomley, A Daemen, A Pexsters, D Timmerman, T Bourne, and A Papageorghiou. The influence of body mass index on first trimester fetal growth. Hum Reprod, In press.

[282] R S Sawhney, B Sharma, L E Humphrey, and M G Brattain. Integrin α_2 and extracellular signal-regulated kinase are functionally linked in highly malignant autocrine transforming growth factor-α-driven colon cancer cells. *J Biol Chem*, 278(22):19861–19869, 2003.

[283] M Schena, D Shalon, R W Davis, and P O Brown. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270:467–470, 1995.

[284] R Schickel, B Boyerinas, S M Park, and M E Peter. MicroRNAs: key players in the immune system, differentiation, tumorigenesis and cell death. *Oncogene*, 27:5959–5974, 2008.

[285] B Schölkopf, K Tsuda, and J-P Vert. *Kernel methods in computational biology*. MIT Press, United States, 2004.

[286] S R Schwarze, J Luo, W B Isaacs, and D F Jarrard. Modulation of CXCL14 (BRAK) expression in prostate cancer. *Prostate*, 64(1):67–74, 2005.

[287] P Sethupathy, B Corda, and A G Hatzigeorgiou. TarBase: a comprehensive database of experimentally supported animal microRNA targets. *RNA*, 12:192–197, 2006.

[288] S Shah, W L Lam, R T Ng, and K P Murphy. Modeling recurrent DNA copy number alterations in array CGH data. *Bioinformatics*, 23:i450–i458, 2007.

[289] S Shah, X Xuan, R J DeLeeuw, M Khojasteh, W L Lam, R T Ng, and K P Murphy. Integrating copy number polymorphisms into array CGH analysis using a robust HMM. *Bioinformatics*, 22(14):e431–e439, 2006.

[290] D S Shames, J D Minna, and A F Gazdar. DNA methylation in health, disease, and cancer. *Curr Mol Med*, 7:85–102, 2007.

[291] D S Shames, J D Minna, and A F Gazdar. Methods for detecting DNA methylation in tumors: from bench to bedside. *Cancer Lett*, 251:187–198, 2007.

[292] J Shawe-Taylor and N Cristianini. *Kernel methods for pattern analysis.* Cambridge University Press, Cambridge, 2004.

[293] K Shedden, J M Taylor, S A Enkemann, M S Tsao, T J Yeatman, W L Gerald, S Eschrich, I Jurisica, T J Giordano, D E Misek, A C Chang, C Q Zhu, D Strumpf, S Hanash, F A Shepherd, K Ding, L Seymour, K Naoki, N Pennell, B Weir, R Verhaak, C Ladd-Acosta, T Golub, and *et al.* Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. *Nat Med*, 14(8):822–827, 2008.

[294] R Shen, D Ghosh, and A M Chinnaiyan. Prognostic meta-signature of breast cancer developed by two-stage mixture modeling of microarray data. *BMC Genomics*, 5:94, 2004.

[295] J Shendure. The beginning of the end for microarrays? *Nat Methods*, 5:585–587, 2008.

[296] J Shendure, G J Porreca, N B Reppas, X Lin, J P McCutcheon, A M Rosenbaum, M D Wang, K Zhang, R D Mitra, and G M Church. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science*, 309:1728–1732, 2005.

[297] A Shlien and D Malkin. Copy number variations and cancer. *Genome Med*, 1:62, 2009.

[298] M Signoretto, A Daemen, C Savorgnan, and J A K Suykens. Variable selection and grouping with multiple graph priors. In *2nd NIPS Workshop on Optim for Mach Learn (OPT)*, 2009.

[299] C J Sigrist, L Cerutti, E de Castro, P S Langendijk-Genevaux, V Bulliard, A Bairoch, and N Hulo. PROSITE, a protein domain database for functional characterization and annotation. *Nucleic Acids Res*, 38:D161–D166, 2010.

[300] R H Silverman. Implications for RNase L in prostate cancer biology. *Biochemistry*, 42:1805–1812, 2003.

[301] C Sima and E R Dougherty. The peaking phenomenon in the presence of feature selection. *Pattern Recognit Lett*, 29:1667–1674, 2008.

[302] R Simon. Roadmap for developing and validating therapeutically relevant genomic classifiers. *J Clin Oncol*, 23:7332–7341, 2005.

[303] R Simon, M K Radmacher, K Dobbin, and L M McShane. Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *J Natl Cancer Inst*, 95:14–18, 2003.

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

[305] L T Smith, G A Otterson, and C Plass. Unraveling the epigenetic code of cancer for therapy. *Trends Genet*, 23(9):449–456, 2007.

[306] A M Snijders, B L Schmidt, J Fridlyand, N Dekker, D Pinkel, R C K Jordan, and D G Albertson. Rare amplicons implicate frequent deregulation of cell fate specification pathways in oral squamous cell carcinoma. *Oncogene*, 24:4232–4242, 2005.

[307] X Solé, N Bonifaci, N López-Bigas, A Berenguer, P Hernández, O Reina, C A Maxwell, H Aguilar, A Urruticoechea, S de Sanjosé, F Comellas, G Capellá, V Moreno, and M A Pujana. Biological convergence of cancer signatures. *PLoS ONE*, 4(2):e4544, 2009.

[308] R L Somorjai, B Dolenko, and R Baumgartner. Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats, cautions. *Bioinformatics*, 19(12):1484–1491, 2003.

[309] C Sotiriou and M J Piccart. Taking gene-expression profiling to the clinic: when will molecular signatures become relevant to patient care? *Nat Rev Cancer*, 7:545–553, 2007.

[310] C Sotiriou, P Wirapati, S Loi, A Harris, S Fox, J Smeds, H Nordgren, P Farmer, V Praz, B Haibe-Kains, C Desmedt, D Larsimont, F Cardoso, H Peterse, D Nuyten, M Buyse, M J van de Vijver, J Bergh, M Piccart, and M Delorenzi. Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *J Natl Cancer Inst*, 98:262–272, 2006.

[311] J A Sparano. TAILORx: trial assigning individualized options for treatment (Rx). *Clin Breast Cancer*, 7(4):347–350, 2006.

[312] J A Sparano and S Paik. Development of the 21-gene assay and its application in clinical practice and clinical trials. *J Clin Oncol*, 26(5):721–728, 2008.

[313] L M Starita and J D Parvin. The multiple nuclear functions of BRCA1: trancription, ubiquitination and DNA repair. *Curr Opin Cell Biol*, 15(3):345–350, 2003.

[314] C Straehle, F Cardoso, E Azambuja, S Dolci, L Meirsman, K Vantongelen, M Ignatiadis, C Sotiriou, and M J Piccart-Gebhart. Better translation from bench to bedside: breakthroughs in the individualized treatment of cancer. *Crit Care Med*, 37(1):S22–S29, 2009.

[315] R L Strausberg, S Levy, and Y H Rogers. Emerging DNA sequencing technologies for human genomic medicine. *Drug Discov Today*, 13:569–577, 2008.

[316] A Subramanian, P Tamayo, V K Mootha, S Mukherjee, B L Ebert, M A Gillette, A Paulovich, S L Pomeroy, T R Golub, E S Lander, and J P Mesirov. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*, 102(43):15545–15550, 2005.

[317] J A K Suykens, J De Brabanter, L Lukas, and J Vandewalle. Weighted Least Squares Support Vector Machines: robustness and sparse approximation. *Neurocomputing*, 48:85–105, 2002.

[318] J A K Suykens, T Van Gestel, J De Brabanter, B De Moor, and J Vandewalle. *Least Squares Support Vector Machines*. World Scientific, Singapore, 2002.

[319] J A K Suykens and J Vandewalle. Least Squares Support Vector Machine classifiers. *Neural Processing Letters*, 9:293–300, 1999.

[320] A Talmor, A Daemen, E Murdoch, H Missfelder-Lobos, D Timmerman, T Bourne, D A Giussani, and C C Lees. The relationship between Doppler pulsatility index, abdominal circumference and growth in severe early onset human fetal growth restriction. Submitted.

[321] I W Taylor, R Linding, D Warde-Farley, Y Liu, C Pesquita, D Faria, S Bull, T Pawson, Q Morris, and J L Wrana. Dynamic modularity in protein interaction networks predicts breast cancer outcome. *Nat Biotech*, 27:199–204, 2009.

[322] R J Tibshirani and B Efron. Pre-validation and inference in microarrays. *Stat Appl Genet Mol Biol*, 1(1):1–18, 2002.

[323] D Timmerman. The use of mathematical models to evaluate pelvic masses: can they beat an expert operator? *Best Pract Res Clin Obstet Gynecol*, 18:91–104, 2004.

[324] D Timmerman, A C Testa, T Bourne, E Ferrazzi, L Ameye, M Konstantinovic, B Van Calster, W Collins, I Vergote, S Van Huffel, and L Valentin. Logistic regression model to distinguish between the benign and malignant adnexal mass before surgery: a multicenter study by the international ovarian tumor analysis group. *J Clin Oncol*, 23:8794–8801, 2005.

[325] D Timmerman, B Van Calster, A C Testa, S Guerriero, D Fischerova, A A Lissoni, C Van Holsbeke, R Fruscio, A Czekierdowski, D Jurkovic, L Savelli, I Vergote, T Bourne, S Van Huffel, and L Valentin. Ovarian cancer prediction in adnexal masses using ultrasound based logistic regression models: a

temporal and external validation study by the IOTA group. *Ultrasound Obstet Gynecol*, 10.1002/uog.7636, 2010.

[326] N Tomioka, S Oba, M Ohira, A Misra, J Fridlyand, S Ishii, Y Nakamura, E Isogai, T Hirata, Y Yoshida, S Todo, Y Kanedo, D G Albertson, D Pinkel, B G Feuerstein, and A Nakagawara. Novel risk stratification of patients with neuroblastoma by genomic signature, which is independent of molecular signature. *Oncogene*, 27:441–449, 2008.

[327] O Troyanskaya, M Cantor, G Sherlock, P Brown, T Hastie, R Tibshirani, D Botstein, and R B Altman. Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17(6):520–525, 2001.

[328] A Uner, Z Akcali, and D Unsal. Serum levels of soluble E-selectin in colorectal cancer. *Neoplasma*, 51(4):269–274, 2004.

[329] Q N Van and T D Veenstra. How close is the bench to the bedside? Metabolic profiling in cancer research. *Genome Med*, 1:5, 2009.

[330] V Van Belle, K Pelckmans, J A K Suykens, and S Van Huffel. On the use of a clinical kernel in survival analysis. In *Eur Symp on Artif Neural Netw, Comput Intell and Mach Learn*, 2010.

[331] B Van Calster, D Timmerman, C Lu, J A K Suykens, L Valentin, C Van Holsbeke, F Amant, I Vergote, and S Van Huffel. Preoperative diagnosis of ovarian tumors using Bayesian kernel-based methods. *Ultrasound Obstet Gynecol*, 29:496–504, 2007.

[332] 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 Rutgers, S Friend, and R Bernards. A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med*, 347(25):1999–2009, 2002.

[333] T Van den Bosch, G Betsas, D Van Schoubroeck, A Daemen, V Vandenbroucke, A Cornelis, B De Moor, J Deprest, and D Timmerman. Gel infusion sonography in the evaluation of the uterine cavity. *Ultrasound Obstet Gynecol*, 34(6):711–714, 2009.

[334] T Van den Bosch, A Daemen, O Gevaert, B De Moor, and D Timmerman. Building decision trees for diagnosing intracavitary uterine pathology. *Facts, Views & Opinions in ObGyn*, 1(3):182–188, 2009.

[335] T Van den Bosch, A Daemen, D Van Schoubroeck, N Pochet, B De Moor, and D Timmerman. Occurrence and outcome of residual trophoblastic tissue: a prospective study. *J Ultrasound Med*, 27(3):357–361, 2008.

[336] T Van den Bosch, V Vandenbroucke, A Daemen, E Domali, D Van Schoubroeck, B De Moor, J Deprest, and D Timmerman. Removal of focal intracavity lesions results in cessation of abnormal uterine bleeding in the vast majority of women. *Ultrasound Obstet Gynecol*, 33(5):612–613, 2009.

[337] T Van den Bosch, J Verguts, A Daemen, O Gevaert, E Domali, P Claerhout, V Vandenbroucke, B De Moor, J Deprest, and D Timmerman. Pain experienced during transvaginal ultrasound, saline contrast sonohysterography, hysteroscopy and office sampling: a comparative study. *Ultrasound Obstet Gynecol*, 31(3):346–351, 2008.

[338] T Van den Bosch, J Verguts, A Daemen, O Gevaert, E Domali, P Claerhout, V Vandenbroucke, B De Moor, J Deprest, and D Timmerman. Reply to re: Pain experienced during transvaginal ultrasound, saline contrast sonohysterography, hysteroscopy and office sampling: a comparative study. *Ultrasound Obstet Gynecol*, 32(1):118–119, 2008.

[339] T Van Gestel, J A K Suykens, B Baesens, S Viaene, J Vanthienen, G Dedene, B De Moor, and J Vandewalle. Benchmarking least squares support vector machine classifiers. *Mach Learn*, 54:5–32, 2004.

[340] T Van Gorp, I Cadron, E Despierre, A Daemen, K Leunen, F Amant, D Timmerman, B De Moor, and I Vergote. HE4 and CA125 as a diagnostic test in ovarian cancer: validation of the risk of ovarian malignancy algorithm (ROMA) in an independent prospective study. Submitted.

[341] C Van Holsbeke, A Daemen, J Yazbek, T K Holland, T Bourne, T Mesens, L Lannoo, A-S Boes, A Joos, A Van De Vijver, N Roggen, B De Moor, E de Jonge, A C Testa, L Valentin, D Jurkovic, and D Timmerman. Ultrasound experience substantially impacts on diagnostic performance and confidence when adnexal masses are classified using pattern recognition. *Gynecol Obstet Invest*, 69(3):160–168, 2010.

[342] C Van Holsbeke, A Daemen, J Yazbek, T K Holland, T Bourne, T Mesens, L Lannoo, B De Moor, E de Jonge, A C Testa, L Valentin, D Jurkovic, and D Timmerman. Ultrasound methods to distinguish between malignant and benign adnexal masses in the hands of examiners with different levels of experience. *Ultrasound Obstet Gynecol*, 34(4):454–461, 2009.

[343] C Van Holsbeke, B Van Calster, A C Testa, E Domali, C Lu, S Van Huffel, L Valentin, and D Timmerman. Prospective internal validation of mathematical models to predict malignancy in adnexal masses: results from the international ovarian tumor analysis study. *Clin Cancer Res*, 15:684–691, 2009.

[344] C Van Holsbeke, B Van Calster, L Valentin, A C Testa, E Ferrazzi, I Dimou, C Lu, P Moerman, S Van Huffel, I Vergote, and D Timmerman. External validation of mathematical models to distinguish between benign and malignant adnexal tumors: a multicenter study by the international ovarian tumor analysis group. *Clin Cancer Res*, 13:4440–4447, 2007.

[345] C Van Holsbeke, J Yazbek, T K Holland, A Daemen, B De Moor, A C Testa, L Valentin, D Jurkovic, and D Timmerman. Real-time ultrasound vs. evaluation of static images in the preoperative evaluation of adnexal masses. *Ultrasound Obstet Gynecol*, 32(6):828–831, 2008.

[346] H Van Malenstein, O Gevaert, L Libbrecht, A Daemen, J Allemeersch, F Nevens, D Cassiman, B De Moor, C Verslype, and J van Pelt. A 7-gene set associated with chronic hypoxia of universal prognostic importance in hepatocellular carcinoma. Submitted.

[347] L J van 't Veer, H Dai, M J van De Vijver, Y D He, A A M Hart, M Mao, H L Peterse, K van der Kooy, M J Marton, A T Witteveen, G J Schreiber, R M Kerkhoven, C Roberts, P S Linsley, R Bernards, and S H Friend. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415:530–536, 2002.

[348] V Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.

[349] J C Venter, M D Adams, E W Myers, P W Li, R J Mural, G G Sutton, H O Smith, M Yandell, C A Evans, R A Holt, J D Gocayne, P Amanatides, R M Ballew, D H Huson, J R Wortman, Q Zhang, C D Kodira, X H Zheng, L Chen, M Skupski, G Subramanian, P D Thomas, J Zhang, and *et al.* The sequence of the human genome. *Science*, 291:1304–1351, 2001.

[350] J-P Vert and M Kanehisa. Extracting active pathways from gene expression data. *Bioinformatics*, 19:ii238–ii244, 2003.

[351] M Vidal. Interactome modeling. *FEBS Letters*, 579:1834–1838, 2005.

[352] H T Viet, D Wågsäter, A Hugander, and J Dimberg. Interleukin-1 receptor antagonist gene polymorphism in human colorectal cancer. *Oncol Rep*, 14(4):915–918, 2005.

[353] B Vogenstein and K W Kinzler. Cancer genes and the pathways they control. *Nature Med*, 10(8):789–799, 2004.

[354] H M Wain, E A Bruford, R C Lovering, M J Lush, M W Wright, and S Povey. Guidelines for human gene nomenclature. *Genomics*, 79(4):464–470, 2002.

[355] H Wainer. Estimating coefficients in linear models: it don't make no nevermind. *Psychological Bulletin*, 83(2):213–217, 1976.

[356] Y Wan, N Wu, Z Wang, X Ju, J Zhu, Y Liu, J Tang, and Y Huang. Relationship between tissue factor expression and hepatic metastasis and prognosis in rectal cancer. *Zhonghua Zhong Liu Za Zhi (Ch J Oncol)*, 24(4):378–380, 2002.

[357] L Wang, R A Baiocchi, S Pal, G Mosialos, M Caligiuri, and S Sif. The BRG1- and hBRM-associated factor BAF57 induces apoptosis by stimulating expression of the cylindromatosis tumor suppressor gene. *Mol Cell Biol*, 25(18):7953–7965, 2005.

[358] S M Wang, L L Ooi, and K M Hui. Identification and validation of a novel gene signature associated with the recurrence of human hepatocellular carcinoma. *Clin Cancer Res*, 13(21):6275–6283, 2007.

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

[360] K M Waters, J G Pounds, and B D Thrall. Data merging for integrated microarray and proteomic analysis. *Brief Funct Genomic Proteomic*, 5(4):261–272, 2006.

[361] D A Wheeler, M Srinivasan, M Egholm, Y Shen, L Chen, A McGuire, W He, Y J Chen, V Makhijani, G T Roth, X Gomes, K Tartaro, F Niazi, C L Turcotte, G P Irzyk, J R Lupski, C Chinault, X Z Song, Y Liu, Y Yuan, L Nazareth, X Qin, D M Muzny, M Margulies, G M Weinstock, R A Gibbs, and J M Rothberg. The complete sequence of an individual by massively parallel DNA sequencing. *Nature*, 452:872–876, 2008.

[362] J M D Wheeler, B F Warren, N J Mortensen, N Ekanyaka, H Kulacoglu, A C Jones, B D George, and M G W Kettlewell. Quantification of histologic regression of rectal cancer after irradiation. *Dis Colon Rectum*, 45:1051–1056, 2002.

[363] L C Whelan, K A R Power, D T McDowell, J Kennedy, and W M Gallagher. Applications of SELDI-MS technology in oncology. *J Cell Mol Med*, 12(5a):1535–1547, 2008.

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

[365] D Witten and R Tibshirani. A comparison of fold-change and the t-statistic for microarray data analysis. Posted at http://www-stat.stanford.edu/~tibs/ftp/FCTComparison.pdf.

[366] I Wolf, N Ben-Baruch, R Shapira-Frommer, S Rizel, H Goldberg, N Yaal-Hahoshen, B Klein, D B Geffen, and B Kaufman. Association between standard clinical and pathologic characteristics and the 21-gene recurrence score in breast cancer patients: a population-based study. *Cancer*, 112:731–736, 2008.

[367] C D Wolfgang, M Essand, B Lee, and I Pastan. T-cell receptor gamma chain alternate reading frame protein (TARP) expression in prostate cancer cells leads to an increased growth rate and induction of caveolins and amphiregulin. *Cancer Res*, 61(22):8122–8126, 2001.

[368] H G Woo, E S Park, J-S Lee, Y-H Lee, T Ishikawa, Y J Kim, and S S Thorgeirsson. Identification of potential driver genes in human liver carcinoma by genomewide screening. *Cancer Res*, 69(9):4059–4066, 2009.

[369] J C Wyatt. Decision support systems. *J R Soc Med*, 93:629–633, 2000.

[370] J C Wyatt and D G Altman. Prognostic models: clinically useful or quickly forgotten? *Br Med J*, 311:1539–1541, 1995.

[371] M Xiong, X Fang, and J Zhao. Biomarker identification by feature wrappers. *Genome Res*, 11:1878–1887, 2001.

[372] Y H Yang, M J Buckley, S Dudoit, and T P Speed. Comparison of methods for image analysis on cDNA microarray data. *J Comput Graphical Stat*, 11(1):108–136, 2002.

[373] Y H Yang, Y Xiao, and M R Segal. Identifying differentially expressed genes from microarray experiments via statistic synthesis. *Bioinformatics*, 21(7):1084–1093, 2005.

[374] T J Yeatman. The future of cancer management: translating the genome, transcriptome, and proteome. *Ann Surg Oncol*, 10(1):7–14, 2003.

[375] E M Youssef, X q Chen, E Higuchi, Y Kondo, G Garcia-Manero, R Lotan, and J P J Issa. Hypermethylation and silencing of the putative tumor suppressor *Tazarotene-induced gene 1* in human cancers. *Cancer Res*, 64:2411–2417, 2004.

[376] J X Yu, A M Sieuwerts, Y Zhang, J W M Martens, M Smid, J G Klijn, Y Wang, and J A Foekens. Pathway analysis of gene signatures predicting metastasis of node-negative primary breast cancer. *BMC Cancer*, 7:182, 2007.

[377] S Yu. *Kernel-based data fusion for machine learning: methods and applications in bioinformatics and text mining*. PhD thesis, Katholieke Universiteit Leuven, Department of Electrical Engineering, 2009.

[378] S Yu, T Falck, A Daemen, L-C Tranchevent, J A K Suykens, B De Moor, and Y Moreau. L2 norm multiple kernel learning and its applications in biomedical data fusion. Submitted.

[379] B Zhang, D Schmoyer, S Kirov, and J Snoddy. GOTree Machine (GOTM): a web-based platform for interpreting sets of interesting genes using Gene Ontology hierarchies. *BMC Bioinformatics*, 5:16, 2004.

[380] L Zhang, S Volinia, T Bonome, G A Calin, J Greshock, N Yang, C G Liu, A Giannakakis, P Alexiou, K Hasegawa, C N Johnstone, M S Megraw, S Adams, H Lassus, J Huang, S Kaur, S Liang, P Sethupathy, A Leminen, V A Simossis, R Sandaltzopoulos, Y Naomoto, D Katsaros, and *et al.* Genomic and epigenetic alterations deregulate microRNA expression in human epithelial ovarian cancer. *Proc Natl Acad Sci U S A*, 105:7004–7009, 2008.

[381] W Zhang, D J Park, B Lu, D Y Yang, M Gordon, S Groshen, J Yun, O A Press, D Vallbohmer, K Rhodes, and H J Lenz. Epidermal growth factor receptor gene polymorphisms predict pelvic recurrence in patients with rectal cancer treated with chemoradiation. *Clin Cancer Res*, 11:600–605, 2005.

[382] Y Zhang, M Lai, B Lv, X Gu, H Wang, Y Zhu, Y Zhu, L Shao, and G Wang. Overexpression of Reg IV in colorectal adenoma. *Cancer Lett*, 200(1):69–76, 2003.

[383] X Zhao, C Li, J G Paez, K Chin, P A Janne, T H Chen, L Girard, J Minna, D Christiani, C Leo, J W Gray, W R Sellers, and M Meyerson. An integrated view of copy number and allelic alterations in the cancer genome using single nucleotide polymorphism arrays. *Cancer Res*, 64:3060–3071, 2004.

[384] F Zinzindohoué, T Lecomte, J M Ferraz, A M Houllier, P H Cugnenc, A Berger, H Blons, and P Laurent-Puig. Prognostic significance of MMP-1 and MMP-3 functional promoter polymorphisms in colorectal cancer. *Clin Cancer Res*, 11(2):594–599, 2005.

Curriculum vitae

Anneleen Daemen was born on July 9, 1983 in Genk, Belgium. In 2006, she obtained the master of science for electrical engineering in electronics option biomedical techniques, summa cum laude from the Katholieke Universiteit Leuven, Belgium. She was laureate biomedical society 2006 and laureate of the KVIV Engineersprice 2007 for her thesis "Physiological modeling of the glucoregulatory system of ICU-patients". Parallel to this, she started music studies option piano at the Lemmensinstitute in Leuven, Belgium and obtained the Bachelor's degree in 2007.

In October 2006, she started her PhD (funded by Scientific Research - Flanders) in the bioinformatics group at the Electrical Engineering Department in Leuven under the supervision of Prof. Dr. Ir. Bart De Moor and Prof. Dr. Dr. Dirk Timmerman. Her research interests include the use of kernel methods for data integration to improve clinical decision support for cancer. In 2008, she received the best student research paper award at the 12th International Conference on Knowledge-Based Engineering Systems for the paper "Classification of sporadic and BRCA1 ovarian cancer based on a genome-wide study of copy number variations".

In September 2010, she will start as postdoctoral BAEF fellow (Belgian American Educational Foundation) in the lab of Prof. Dr. Paul Spellman and Prof. Dr. Joe Gray at Lawrence Berkeley National Laboratory.

Publications by the author

Papers in International Journals

A. Daemen, O. Gevaert, F. Ojeda, A. Debucquoy, J.A.K. Suykens, C. Sempoux, J.-P. Machiels, K. Haustermans, and B. De Moor. A kernel-based integration of genome-wide data for clinical decision support. *Genome Med*, 1(4):39, 2009.

A. Daemen, M. Signoretto, O. Gevaert, J.A.K. Suykens, and B. De Moor. Improved microarray-based decision support with graph encoded interactome data. *PLoS ONE*, 5(4), e10225, 2010.

T. Van den Bosch, J. Verguts, **A. Daemen**, O. Gevaert, E. Domali, P. Claerhout, V. Vandenbroucke, B. De Moor, J. Deprest, and D. Timmerman. Pain experienced during transvaginal ultrasound, saline contrast sonohysterography, hysteroscopy and office sampling: a comparative study. *Ultrasound Obstet Gynecol*, 31(3), 346-351, 2008. (IF 2.690)

T. Van den Bosch, **A. Daemen**, D. Van Schoubroeck, N. Pochet, B. De Moor, and D. Timmerman. Occurrence and outcome of residual trophoblastic tissue: a prospective study. *J Ultrasound Med*, 27(3), 357-361, 2008. (IF 1.087)

C. Van Holsbeke, J. Yazbek, T.K. Holland, **A. Daemen**, B. De Moor, A.C. Testa, L. Valentin, D. Jurkovic, and D. Timmerman. Real-time ultrasound vs. evaluation of static images in the preoperative evaluation of adnexal masses. *Ultrasound Obstet Gynecol*, 32(6), 828-831, 2008. (IF 2.690)

T. Van den Bosch, J. Verguts, **A. Daemen**, O. Gevaert, E. Domali, P. Claerhout, V. Vandenbroucke, B. De Moor, J. Deprest, and D. Timmerman. Reply to Re: Pain experienced during transvaginal ultrasound, saline contrast sonohysterography, hysteroscopy and office sampling: a comparative study. *Ultrasound Obstet Gynecol*, 32(1), 118-119, 2008. (IF 2.690)

T. Van den Bosch, V. Vandenbroucke, **A. Daemen**, E. Domali, D. Van Schoubroeck, B. De Moor, J. Deprest, and D. Timmerman. Removal of focal

intracavity lesions results in cessation of abnormal uterine bleeding in the vast majority of women. *Ultrasound Obstet Gynecol*, 33(5), 612-613, 2009. (IF 2.690)

E. Kirk, **A. Daemen**, A.T. Papageorghiou, C. Bottomley, G. Condous, B. De Moor, D. Timmerman, and T. Bourne. Why are some ectopic pregnancies characterized as pregnancies of unknown location at the initial transvaginal ultrasound examination. *Acta Obstet Gynecol Scand*, 87(11), 1150-1154, 2008. (IF 1.356)

C. Bottomley, **A. Daemen**, F. Mukri, A.T. Papageorghiou, E. Kirk, A. Pexsters, B. De Moor, D. Timmerman, and T. Bourne. Functional linear discriminant analysis: a new longitudinal approach to the assessment of embryonic growth. *Hum Reprod*, 24(2), 278-283, 2009. (IF 3.773)

C. Bottomley, **A. Daemen**, F. Mukri, A.T. Papageorghiou, E. Kirk, A. Pexsters, B. De Moor, D. Timmerman, and T. Bourne. Assessing first trimester growth: the influence of ethnic background and maternal age. *Hum Reprod*, 24(2), 284-290, 2009. (IF 3.773)

A. Debucquoy, K. Haustermans, **A. Daemen**, S. Aydin, L. Libbrecht, O. Gevaert, S. Tejpar, W.H. McBride, F. Penninckx, P. Scalliet, C. Stroh, S. Vlassak, C. Sempoux, and J.-P.Machiels. Molecular response to cetuximab and efficacy of preoperative cetuximab-based chemoradiation in rectal cancer. *J Clin Oncol*, 27(17), 2751-2757, 2009. (IF 17.157)

I. Cadron, T. Van Gorp, F. Amant, I. Vergote, P. Moerman, E. Waelkens, **A. Daemen**, R. Van de Plas, B. De Moor, and R. Zeillinger. The use of laser microdissection and SELDI-TOF MS in ovarian cancer tissue to identify protein profiles. *Anticancer Res*, 29(4), 1039-1046, 2009. (IF 1.390)

T. Van den Bosch, G. Betsas, D. Van Schoubroeck, **A. Daemen**, V. Vandenbroucke, A. Cornelis, B. De Moor, J. Deprest, and D. Timmerman. Gel infusion sonography in the evaluation of the uterine cavity. *Ultrasound Obstet Gynecol*, 34(6), 711-714, 2009. (IF 2.690)

C. Van Holsbeke, **A. Daemen**, J. Yazbek, T.K. Holland, T. Bourne, T. Mesens, L. Lannoo, B. De Moor, E. de Jonge, A.C. Testa, L. Valentin, D. Jurkovic, and D. Timmerman. Ultrasound methods to distinguish between malignant and benign adnexal masses in the hands of examiners with different levels of experience. *Ultrasound Obstet Gynecol*, 34(4), 454-461, 2009. (IF 2.690)

K. Leunen, O. Gevaert, **A. Daemen**, V. Vanspauwen, G. Michils, B. De Moor, P. Moerman, I. Vergote, and E. Legius. Recurrent copy number alterations in BRCA1-mutated ovarian tumors alter biological pathways. *Hum Mut*, 30(12), 1693-1702, 2009. (IF 7.033)

L.A.M. Gravendeel, M.C.M. Kouwenhoven, O. Gevaert, J.J. de Rooi, A.P. Stubbs,

J.E. Duijm, **A. Daemen**, F.E. Bleeker, L.B.C. Bralten, N.K. Kloosterhof, B. De Moor, P.H. Eilers, P.J. van der Spek, J.M. Kros, P.A.E. Sillevis Smitt, M.J. van den Bent, and P.J. French. Intrinsic gene expression profiles of gliomas are a better predictor of survival than histology. *Cancer Res*, 69(23), 9065-9072, 2009. (IF 7.514)

O. Gevaert, **A. Daemen**, B. De Moor, and L. Libbrecht. A taxonomy of epithelial human cancer and their metastases. *BMC Med Genomics*, 2, 69, 2009.

T. Van den Bosch, **A. Daemen**, O. Gevaert, B. De Moor, and D. Timmerman. Building decision trees for diagnosing intracavitary uterine pathology. *Facts, Views & Opinions in ObGyn*, 1(3), 182-188, 2009.

C. Van Holsbeke, **A. Daemen**, J. Yazbek, T.K. Holland, T. Bourne, T. Mesens, L. Lannoo, A.-S. Boes, A. Joos, A. Van De Vijver, N. Roggen, B. De Moor, E. de Jonge, A.C. Testa, L. Valentin, D. Jurkovic, and D. Timmerman. Ultrasound experience substantially impacts on diagnostic performance and confidence when adnexal masses are classified using pattern recognition. *Gynecol Obstet Invest*, 69(3), 160-168, 2010. (IF 1.417)

A. Pexsters, **A. Daemen**, C. Bottomley, D. Van Schoubroeck, L. De Catte, B. De Moor, T. D'Hooghe, C. Lees, D. Timmerman, and T. Bourne. A new crown-rump length size curve based on over 3500 pregnancies. *Ultrasound Obstet Gynecol*, In press, 2010. (IF 2.690)

I. Sarris, C. Bottomley, **A. Daemen**, A. Pexsters, D. Timmerman, T. Bourne, and A. Papageorghiou. The influence of body mass index on first trimester fetal growth. *Hum Reprod*, In press, 2010. (IF 3.773)

Book Chapters

A. Daemen, O. Gevaert, K. Leunen, V. Vanspauwen, G. Michils, E. Legius, I. Vergote, and B. De Moor. A genome-wide computational study of copy number variations: an example on ovarian cancer. Chapter 9 of *Investigating human cancer with computational intelligence techniques* (A. Vellido, and P. Lisboa eds), In *KES Recent Research Results series*. Future Technology Press, Shoreham-by-sea, UK, 2009, 107-118.

Papers Submitted to International Journals

A. Daemen, O. Gevaert, F. Ojeda, J.A.K. Suykens, and B. De Moor. HI-DID-IT, a HIgh-DImensional Data Integration Toolbox for clinical applications. Submitted

to *BMC Bioinformatics.* (IF 3.781)

A. Daemen, D. Timmerman, T. Van den Bosch, C. Bottomley, E. Kirk, C. Van Holsbeke, L. Valentin, T. Bourne, and B. De Moor. Improved modeling of clinical data with kernel methods. Submitted to *Artif Intell Med.* (IF 1.960)

A. Daemen, L. Valentin, R. Fruscio, C. Van Holsbeke, G.B. Melis, S. Guerriero, A. Czekierdowski, D. Jurkovic, W. Ombelet, A. Rossi, I. Vergote, T. Bourne, B. De Moor, and D. Timmerman. Improving the preoperative classification of adnexal masses as benign or malignant with second stage tests when the basic test is subjective evaluation of ultrasound findings or a mathematical formula for the calculation of risk of malignancy. Revised manuscript submitted to *Ultrasound Obstet Gynecol.* (IF 2.690)

A. Daemen, D. Jurkovic, C. Van Holsbeke, S. Guerriero, A.C. Testa, A. Czekierdowski, R. Fruscio, D. Paladini, P. Neven, A. Rossi, T. Bourne, B. De Moor, and D. Timmerman. The effect of cancer prevalence on the use of different cut-off levels and the performance of mathematical models to distinguish benign from malignant adnexal masses. Submitted to *Ultrasound Obstet Gynecol.* (IF 2.690)

S. Yu, T. Falck, **A. Daemen**, L.-C. Tranchevent, J.A.K. Suykens, B. De Moor, and Y. Moreau. L2 norm multiple kernel learning and its applications in biomedical data fusion. Revised manuscript submitted to *BMC Bioinformatics.* (IF 3.781)

T. Van den Bosch, D. Van Schoubroeck, **A. Daemen**, E. Domali, V. Vandenbroucke, B. De Moor, J. Deprest, and D. Timmerman D. Lidocaine does not reduce pain perception during gel instillation sonography (GIS) or subsequent office hysteroscopy – results of a randomized trial. Revised manuscript submitted to *Hum Reprod.* (IF 3.773)

O. Habayeb, **A. Daemen**, D. Timmerman, B. De Moor, G.A. Hackett, T. Bourne, and C.C. Lees. The relationship between first trimester fetal growth, pregnancy associated plasma protein A levels and birthweight. Submitted to *Hum Reprod.* (IF 3.773)

H. Van Malenstein, O. Gevaert, L. Libbrecht, **A. Daemen**, J. Allemeersch, F. Nevens, D. Cassiman, B. De Moor, C. Verslype, and J. van Pelt. A 7-gene set associated with chronic hypoxia of universal prognostic importance in hepatocellular carcinoma. Revised manuscript submitted to *Clin Cancer Res.* (IF 6.488)

A. Talmor, **A. Daemen**, E. Murdoch, H. Missfelder-Lobos, D. Timmerman, T. Bourne, D.A. Giussani, and C.C. Lees. The relationship between Doppler pulsatility index, abdominal circumference and growth in severe early onset human fetal growth restriction. Submitted to *J Clin Invest.* (IF 16.559)

T. Van Gorp, I. Cadron, E. Despierre, **A. Daemen**, K. Leunen, F. Amant, D. Timmerman, B. De Moor, and I. Vergote. HE4 and CA125 as a diagnostic test in ovarian cancer: validation of the risk of ovarian malignancy algorithm (ROMA) in an independent prospective study. Submitted to *Cancer*. (IF 5.238)

Papers in Preparation

A. Pexsters, **A. Daemen**, J.P. Frijns, G. Betsas, C. Bottomley, D. Van Schoubroeck, B. De Moor, T. D'Hooghe, C. Lees, D. Timmerman, and T. Bourne. Crown rump length in genetically abnormal pregnancies compared to a reference CRL size curve of normal pregnancies. To be submitted to *Ultrasound Obstet Gynecol.* (IF 2.690)

A. Smeets, **A. Daemen**, I. Van den Bempt, O. Gevaert, B. Claes, H. Wildiers, R. Drijkoningen, P. Van Hummelen, D. Lambrechts, P. Neven, B. De Moor, T. Van Dorpe, R. Paridaens, M.R. Christiaens. Prediction of lymph node involvement in breast cancer from primary tumor tissue using gene expression profiling and miRNAs. To be submitted to *Clin Cancer Res.* (IF 6.488)

U. Hussain, **A. Daemen**, H. Missfelder-Lobos, B. De Moor, D. Timmerman, T. Bourne, and C.C. Lees. Umbilical artery PI and fetal abdominal circumference in isolated gastroschisis.

A. Pexsters, **A. Daemen**, C. Bottomley, Y. Abdallah, O. Naji, N. Raine-Fenning, B. De Moor, T. D'Hooghe, D. Timmerman, and T. Bourne. A longitudinal assessment of gestational sac growth in pregnancies subsequently shown to be viable and non-viable.

M. Signoretto, **A. Daemen**, C. Savorgnan, and J.A.K. Suykens. Networks-aided learning for high dimensional sparse models.

Papers in International Conference Proceedings

A. Daemen, O. Gevaert, and B. De Moor B. Integration of clinical and microarray data with kernel methods. *Proceedings of the 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Lyon, France, pp. 5411-5415, Aug. 2007.

A. Daemen, O. Gevaert, T. De Bie, A. Debucquoy, J.-P. Machiels, B. De Moor, and K. Haustermans. Integrating microarray and proteomics data to predict the response on cetuximab in patients with rectal cancer. *Pacific Symposium on Biocomputing (PSB)*, Big Island, Hawaii, pp. 166-177, Jan. 2008.

A. Daemen, O. Gevaert, K. Leunen, V. Vanspauwen, G. Michils, E. Legius, I. Vergote, and B. De Moor. Classification of sporadic and BRCA1 ovarian cancer based on a genome-wide study of copy number variations. *Knowledge-Based Intelligent Information and Engineering Systems* (Lovrek I, Howlett RJ, and Jain LC eds), Lect Notes Comput Sci, vol. 5178, Springer, pp. 165-172, 2008.

A. Daemen, O. Gevaert, K. Leunen, L. Legius, I. Vergote, and B. De Moor. Supervised classification of array CGH data with HMM-based feature selection. *Pacific Symposium on Biocomputing (PSB)*, Big Island, Hawaii, pp. 468-479, Jan. 2009.

A. Daemen, and B. De Moor. Development of a kernel function for clinical data. *Proceedings of the 31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Minnesota, USA, pp. 5913-5917, Sept. 2009.

M. Signoretto, **A. Daemen**, C. Savorgnan, and J.A.K. Suykens. Variable selection and grouping with multiple graph priors. *2nd NIPS Workshop on Optimization for Machine Learning (OPT)*, Whistler, Canada, Dec. 2009.

Conference Abstracts in International Journals

17th World Congress on Ultrasound in Obstetrics and Gynecology, Firenze-Florence, Italy, Oct. 2007

A. Daemen, C. Bottomley, O. Gevaert, B. De Moor, D. Timmerman, and T. Bourne. Predicting early pregnancy loss with Functional Linear Discriminant Analysis (FLDA). *Ultrasound Obstet Gynecol*, 30(4), 469, 2007. (IF 2.690)

T. Van den Bosch, **A. Daemen**, O. Gevaert, and D. Timmerman. Mathematical decision trees versus clinician based algorithms in the diagnosis of endometrial disease. *Ultrasound Obstet Gynecol*, 30(4), 412, 2007. (IF 2.690)

C. Van Holsbeke, J. Yazbek, **A. Daemen**, T. Holland, A.C. Testa, L. Valentin, D. Timmerman, and D. Jurkovic. Real-time ultrasound versus static image evaluation of adnexal masses. *Ultrasound Obstet Gynecol*, 30(4), 429, 2007. (IF 2.690)

J. Yazbek, C. Van Holsbeke, **A. Daemen**, T. Holland, A.C. Testa, L. Valentin, D. Timmerman, and D. Jurkovic. How accurate is ultrasound pattern recognition at predicting the histological diagnosis of an ovarian mass? *Ultrasound Obstet Gynecol*, 30(4), 429, 2007. (IF 2.690)

American Society of Clinical Oncology Annual Meeting, Chicago, USA, June 2008

J.-P. Machiels, A. Debucquoy, O. Gevaert, **A. Daemen**, C. Sempoux, W. McBride, C. Stroh, S. Vlassak, and K. Haustermans. Prediction of pathological response to preoperative chemoradiotherapy with cetuximab in rectal cancer. *J Clin Oncol*, 26(15S), 4095, 2008. (IF 17.157)

24th Annual Meeting of the European Society Human Reproduction and Embryology, Barcelona, Spain, July 2008

A. Daemen, C. Bottomley, F. Mukri, E. Kirk, A.T. Papageorghiou, D. Timmerman, B. De Moor, and T. Bourne. A new statistical method for longitudinal assessment of early fetal growth – prospective testing of a functional linear discriminant analysis model. *Hum Reprod*, 23(1), i43-i45, 2008. (IF 3.773)

C. Bottomley, **A. Daemen**, F. Mukri, E. Kirk, A.T. Papageorghiou, D. Timmerman, B. De Moor, and T. Bourne. Revisiting very early first trimester growth - influence of ethnic background and maternal age using transvaginal ultrasonography (TVS). *Hum Reprod*, 23(1), i43-i45, 2008. (IF 3.773)

18th World Congress on Ultrasound in Obstetrics and Gynecology, Chicago, USA, Aug. 2008

A. Daemen, C. Van Holsbeke, R. Fruscio, S. Guerriero, A. Czekierdowski, L. Valentin, L. Savelli, A.C. Testa, N. Colombo, T. Bourne, I. Vergote, B. De Moor, and D. Timmerman. Prevalence of cancer and optimal cut-off levels for mathematical models to distinguish between benign and malignant adnexal masses. *Ultrasound Obstet Gynecol*, 32(3), 292-293, 2008. (IF 2.690)

A. Daemen, C. Van Holsbeke, O. Gevaert, R. Fruscio, S. Guerriero, A. Czekierdowski, L. Valentin, L. Savelli, A.C. Testa, D. Fischerova, T. Bourne, I. Vergote, B. De Moor, and D. Timmerman. Prospective comparison of one-step and two-step models for the classification of adnexal masses as benign and malignant. *Ultrasound Obstet Gynecol*, 32(3), 368, 2008. (IF 2.690)

C. Bottomley, **A. Daemen**, F. Mukri, A.T. Papageorghiou, E. Kirk, A. Pexsters, B. De Moor, D. Timmerman, and T. Bourne. Growth delay in the first trimester: sensitivity and specificity of using serial measurements of embryonic growth for the prediction of early pregnancy loss. *Ultrasound Obstet Gynecol*, 32(3), 278, 2008. (IF 2.690)

C. Bottomley, **A. Daemen**, F. Mukri, E. Kirk, A.T. Papageorghiou, A. Pexsters, B. De Moor, D. Timmerman, and T. Bourne. Assessing early first trimester growth with ultrasonography: the influence of ethnic background. *Ultrasound Obstet Gynecol*, 32(3), 354, 2008. (IF 2.690)

C. Bottomley, **A. Daemen**, F. Mukri, A.T. Papageorghiou, E. Kirk, A. Pexsters, B. De Moor, D. Timmerman, and T. Bourne. The influence of maternal age on early first trimester growth assessed by ultrasound. *Ultrasound Obstet Gynecol*, 32(3), 355, 2008. (IF 2.690)

A. Pexsters, **A. Daemen**, C. Bottomley, D. Van Schoubroeck, L. De Catte, D. Timmerman, and T. Bourne. Validation of the Robinson growth curve with data of transvaginal ultrasound assessments in early pregnancy from a large Caucasian population. *Ultrasound Obstet Gynecol*, 32(3), 278, 2008. (IF 2.690)

A. Pexsters, **A. Daemen**, J.P. Frijns, C. Bottomley, D. Van Schoubroeck, L. De Catte, D. Timmerman, and T. Bourne. A cross-sectional analysis of embryonic growth in normal and genetically abnormal pregnancies. *Ultrasound Obstet Gynecol*, 32(3), 337, 2008. (IF 2.690)

O. Gevaert, A.C. Testa, **A. Daemen**, C. Van Holsbeke, R. Fruscio, E. Epstein, F.P.G. Leone, A. Czerkierdowski, L. Valentin, L. Savelli, T. Bourne, F. Amant, B. De Moor, and D. Timmerman. Investigation of the performance of mathematical models on small ovarian masses on IOTA phase 1 and 2 data. *Ultrasound Obstet Gynecol*, 32(3), 292, 2008. (IF 2.690)

C. Van Holsbeke, **A. Daemen**, L. Lannoo, T. Mesens, E. de Jonge, L. Valentin, D. Jurkovic, J. Yazbek, T.K. Holland, and D. Timmerman. Pattern recognition by less experienced examiners and use of mathematical models to discriminate between static ultrasound images of benign and malignant adnexal masses. *Ultrasound Obstet Gynecol*, 32(3), 292, 2008. (IF 2.690)

T. Van den Bosch, V. Vandenbroucke, **A. Daemen**, E. Domali, D. Van Schoubroeck, B. De Moor, J. Deprest, and D. Timmerman. Removal of focal intracavity lesions results in cessation of abnormal uterine bleeding in the vast majority of women. *Ultrasound Obstet Gynecol*, 32(3), 258, 2008. (IF 2.690)

E. Kirk, A.T. Papageorghiou, **A. Daemen**, G. Condous, D. Timmerman, and T. Bourne. The appearance and behaviour of ectopic pregnancies in the 'pregnancy of unknown location' population. *Ultrasound Obstet Gynecol*, 32(3), 273, 2008. (IF 2.690)

19th World Congress on Ultrasound in Obstetrics and Gynecology, Hamburg, Germany, Sep. 2009

A. Pexsters, **A. Daemen**, J.P. Frijns, C. Bottomley, D. Van Schoubroeck, L. De Catte, B. De Moor, T. D'Hooghe, D. Timmerman, and T. Bourne. Crown-rump length in genetically abnormal pregnancies compared to a CRL curve for normal pregnancies. *Ultrasound Obstet Gynecol*, 34(S1), 85, 2009. (IF 2.690)

A. Pexsters, **A. Daemen**, C. Bottomley, D. Van Schoubroeck, L. De Catte, B. De Moor, T. D'Hooghe, C. Lees, D. Timmerman, and T. Bourne. A new crown-rump length curve based on over 3500 pregnancies. *Ultrasound Obstet Gynecol*, 34(S1), 11-12, 2009. (IF 2.690)

C. Bottomley, **A. Daemen**, A. Pexsters, E. Kirk, D. Timmerman, A.T. Papageorghiou, and T. Bourne. The influence of maternal demographic features on fetal size in the first trimester. *Ultrasound Obstet Gynecol*, 34(S1), 92, 2009. (IF 2.690)

I. Sarris, **A. Daemen**, C. Bottomley, A. Pexsters, D. Timmerman, T. Bourne, and A.T. Papageorghiou. The influence of body mass index on first trimester fetal growth. *Ultrasound Obstet Gynecol*, 34(S1), 92-93, 2009. (IF 2.690)

O. Habayeb, **A. Daemen**, D. Timmerman, B. De Moor, T. Bourne, G. Hackett, and C. Lees. The relationship between early first trimester PAPP-A and birthweight percentiles. *Ultrasound Obstet Gynecol*, 34(S1), 253, 2009. (IF 2.690)

O. Habayeb, **A. Daemen**, D. Timmerman, B. De Moor, T. Bourne, G. Hackett, and C. Lees. The relationship between fetal growth in the first trimester, PAPP-A and birthweight percentiles. *Ultrasound Obstet Gynecol*, 34(S1), 95, 2009. (IF 2.690)

A. Talmor, **A. Daemen**, E. Murdoch, B. De Moor, H. Missfelder-Lobos, D. Timmerman, T. Bourne, and C. Lees. The relationship between abdominal circumference and umbilical artery Doppler z-scores in severe early onset fetal growth restriction. *Ultrasound Obstet Gynecol*, 34(S1), 151, 2009. (IF 2.690)

A. Talmor, **A. Daemen**, E. Murdoch, H. Missfelder-Lobos, B. De Moor, D. Timmerman, T. Bourne, and C. Lees. The relationship between abdominal circumference and middle cerebral artery Doppler z-scores in severe early onset fetal growth restriction. *Ultrasound Obstet Gynecol*, 34(S1), 257, 2009. (IF 2.690)

Arenberg Doctoral School of Science, Engineering & Technology
Faculty of Engineering
Department of Electrical Engineering (ESAT)
SCD/SISTA - Bioinformatics
Kasteelpark Arenberg 10 box 2446
B-3001 Leuven

