UNIVERSITE LIBRE DE BRUXELLES
Faculte Des Sciences
Departement D'Informatique

Use of Machine Learning in Bioinformatics to Identify Prognostic and Predictive Molecular Signatures in Human Breast Cancer

Promoteur :
M. Gianluca Bontempi
Co-promoteur :
M. Christos Sotiriou

Mémoire présenté
en vue de l'obtention
du DEA en Sciences
par Benjamin Haibe-Kains

Année Académique 2004 - 2005

Contents

Acknowledgments

I would like to thank so many people that it is impossible to find any order for their respective contributions. Let's start :

G. Bontempi who followed me since my licence in computer science and initiated the collaboration with Christos Sotiriou. His help was invaluable for my research and his *risotto alla milanese* has changed my life !

Christos Sotiriou who has allowed me to do my training in his lab and who believes enough in me to keep me. His passion for the research is a great source of motivation. And last but not least, I thank him for his confidence that allows me to continue my research in spite of the huge amount of work in the lab.

All my colleagues of the Microarray Unit at the IJB for their enthusiasm and their efficiency. Christine Desmedt, Francoise Lallemand, Virginie Durbecq and Sherene Loi for their great discussions and their motivation.

All my colleagues of the Machine Learning Group at the ULB. I have passed great moments with them during this first year of research. Special thanks to Yann-Aël Le Borgne to have read with courage the preliminary versions of my thesis.

Raymond Devillers for his careful reading.

My girlfriend, Olivia, to support me even when I live only for my work.

Finally all my teachers who have opened my mind to such interesting research fields.

I hope that my friends and my family have not really suffered from my bad mood during intensive work. Their support was not only valuable but necessary.

Chapter 1

Introduction

Contents

Thanks to the routine use of screening mammograms in developed countries, more and more women are diagnosed with early breast cancer (small tumors and absence of lymph node invasion). However, despite early detection, up to 20 to 30% of these women will relapse and die from their disease. The majority of these deaths are due to distant metastases. Loco-regional treatment (surgery and radiotherapy) are always carried out and a systemic adjuvant treatment (e.g. chemotherapy and/or endocrine therapy) is proposed to all *high-risk* patients to prevent recurrence.

The definition of such a risk is a central problem in clinic and can have two different significations. The risk can have a *prognostic* value which is its power of prediction of survival independently of treatment. On the other hand, the risk can have a *predictive* value which is its power of prediction of survival under treatment.

Currently the risk is defined from several histological criteria established during consensus conferences in Europe and USA [Goldhirsh et al., 1998; Eifel et al., 2001; Goldhirsh et al., 2003] which attempt to define prognostic criteria for breast cancers[1] :

- Invasive/non-invasive breast cancer :
  - Non-invasive (or "in situ") cancers confine themselves to the ducts or lobules and do not spread to the surrounding tissues in the breast or other parts of the body. However they can develop into or increase your risk for invasive cancer.
  - Invasive (or infiltrating) cancers have started to break through normal breast tissue barriers and invade surrounding areas. Much more serious than non-invasive

[1] `http://www.breastcancer.org`

cancers, invasive cancers can spread to other parts of the body through the bloodstream and lymphatic system.

- Number of involved lymph nodes : some breast cancers spread to the lymph nodes under the arm. When the lymph nodes are involved in the cancer, they are called "node positive". When lymph nodes are free of cancer, they are called "node negative". In large medical studies, there appears to be a correlation between the number of involved lymph nodes and the cancer aggressiveness. Knowing how many lymph nodes are affected by cancer can help to select the more aggressive treatment in the adjuvant setting.

- Tumor size : tumors with large tumor size are considered poor prognosis. Currently, the breast tumors are diagnosed earlier and consequently, their size is smaller.

- Tumor rate/grade :

  - Rate of cancer cell growth : the proportion of cancer cells growing and making new cells varies from tumor to tumor and may be helpful in predicting how aggressive a cancer is. If more than 6-10% of the cells are making new cells, the rate of growth is considered unfavorably high.
  - Grade of cancer cell growth : patterns of cell growth are rated on a scale from 1 to 3 (also referred to as low, medium, and high instead of 1, 2 or 3). Calm, well-organized growth with few cells reproducing is considered grade 1. Disorganized, irregular growth patterns in which many cells are in the process of making new cells is called grade 3. The lower the grade, the more favorable the expected outcome. At the same time, the higher the grade, the more vulnerable the cancer is to treatments such as chemotherapy and radiation. Thus, the histological grade in breast cancer provides important prognostic information. However, its inter-observer variability and poor reproducibility, especially for tumours of intermediate grade, has limited its clinical potential. A recent study [Sotiriou et al., 2005] has determined a refinement of the histological grade using gene expression profiling.
  - Dead cells within the tumor : it is tempting to think that the only good cancer cell is a dead cancer cell. However, necrosis (or dead tumor cells) is one of several signs of excessive tumor growth.

- Hormone receptor status : estrogen and progesterone stimulate the growth of normal breast cells as well as some breast cancer cells. If a tumor is estrogen-receptor positive (ER-positive), it is more likely to grow in a high-estrogen environment. ER-negative tumors are usually not affected by the levels of estrogen and progesterone in your body. ER-positive cancers are more likely to respond to anti-estrogen therapies (e.g. TAMOXIFEN, a drug that works by blocking the estrogen receptors on the breast tissue cells and slowing their estrogen-fueled growth).

- Oncogenes : according to the oncogene, it is either the gene amplification, the increasing amount of its protein or its mutation that confer its properties in breast cancer. The over-expression happens when an oncogene (such as HER2/neu, EGFR, and p53) over-expresses itself by making excessive normal or abnormal proteins and receptors. Cancers that result from over-expressed oncogenes tend to be more nasty or belligerent and are more likely to recur than other cancers. They also may respond to different types of treatment than other breast cancers.

- Margins of resection : the term “margins”or “margins of resection”is used to refer to the distance between the tumor and the edge of the tissue taken by surgery. The margins are measured on all six sides: front and back, top and bottom, left and right.

According to these histological criteria, approximately 80% of young patients without lymph node invasion are candidates for adjuvant treatment. It is obvious that these patients are over-treated because 70 to 80% of them will not develop distant metastases without the adjuvant treatment [EBCT Collaborative Group, 1998]. These results highlight the necessity to improve the risk evaluation based on traditional factors.

During last ten years, several prognosis factors (e.g. HER2 and p53 mutations) have been assessed and have been correlated to the prognosis but these genes, taken individually, have only a limited prognostic power. Moreover, intensive research concerns specific markers for treatment response but these markers have only limited predictive power. This is probably due to the molecular complexity and heterogeneity of the tumors. The tumor phenotype is not determined by isolated aberrations but by a combination of anomalies in a genetic context.

Currently, thanks to technological advances in genome sequencing, new tools are available to analyze biological materials at the molecular level. The microarray technology (which will be introduced in Section 3.2) allows to analyze the genetic identity of a specific tissue for the whole genome. In one microarray experiment, the expression of several thousands of genes can be measured from a tumor tissue. This technology can be used to study the molecular make-up of multiple breast tumors to improve the risk evaluation and our understanding of this biological phenomenon.

1.1 Bioinformatics Context

The use of machine learning methods [Mitchell, 1997; Hastie et al., 2001] in the field of bioinformatics is increasing over time. Such methods seem to be good candidates to treat microarray data [Dudoit et al., 2002].

Many problems in genomics are analyzed by machine learning methods. These include cancer prediction, gene finding, protein structures and functions, protein interactions, gene regulation networks, among many other problems.

Here is a definition of machine learning[2] :

> *Machine learning is a field of artificial intelligence related to data mining and statistics. It involves learning from data. The researcher feeds a set of training examples to a computer program that aims to learn the connection between features of the examples and a specified target concept.*

In our problem, the expression values of the genes are the input and the target concept is the survival of the corresponding patients.

An important example of the use of machine learning methods in human breast cancer is the prognosis of node-negative breast cancers using microarray [van’t Veer et al., 2002]. According to a common view, progression from a primary to a metastatic tumor is accompanied

[2] The definition comes from `http://en.wikipedia.org/wiki/Machine_Learning`.

by the sequential acquisition of phenotype changes, thus allowing breast cancer cells to invade, disseminate, and colonize distant sites. Nevertheless, most investigations have revealed that progression is not accompanied by major changes in marker expression or grade [Lacroix and Leclercq, 2004].

These observations suggest that the metastatic signature might already be present in the primary breast tumor, challenging the traditional model of metastasis, which specifies that most primary tumor cells have low metastatic potential, but rare cells within large primary tumors acquire metastatic capacity through somatic mutations.

From that perspective, [van't Veer et al., 2002], applying a machine learning method (supervised learning, see Figure 1.1), sought to identify whether there exists a gene expression signature strongly prognostic of a short interval to distant metastases in primary breast cancer tumors.

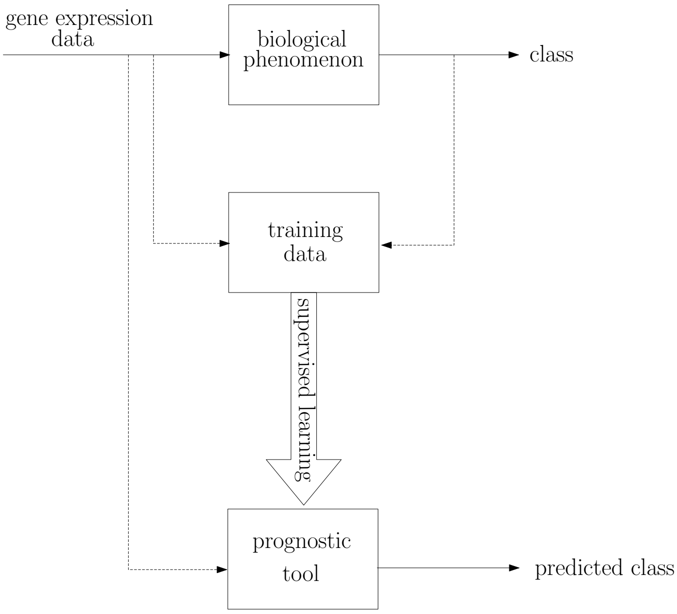


Figure 1.1: Supervised learning method used in microarray classification as in [van't Veer et al., 2002]. The learning method constructs a classifier on the basis of the microarray data (gene expressions) and survival information about the patients (i.e. binary class representing the appearance of distant metastases in the first 5 years of follow-up). We can use this classifier to predict the class of new data (i.e. a tumor tissue from a new patient).

They found 231 genes significantly associated with disease outcome as defined by the presence of distant metastasis at the 5-year mark. They could then subsequently collapse this list into a core set of 70 prognostic markers. Interestingly, the investigators tested the ability of this array-derived prognostic "expression profile" to correctly identify patients who would need adjuvant chemotherapy and compared it to accepted guidelines for treatment of node negative breast cancer (NIH [Eifel et al., 2001] and St. Gallen [Goldhirsh et al., 1998] consensus guidelines). They found that although the expression profile could correctly identify patients who would need adjuvant chemotherapy, it could effectively reduce the fraction of women not needing adjuvant chemotherapy by about 30%. The same group applied this signature to a larger test set of node negative and node positive breast cancer patients (295) from the same institution. This study confirmed that the 70-gene prognosis signature could clearly distinguish patients with excellent 10 year survival from those with a high mortality rate [van de Vijver et al., 2002].

In this thesis, we propose a machine learning methodology (which will be described in Chapter 4). We perform an experimental validation on real microarray data concerning the prediction of treatment resistance for breast cancer patients. The microarray data come from the Microarray Unit of the Institut Jules Bordet.

1.1.1 Treatment Resistance in Breast Cancer

One of the most important advances in the treatment of breast cancer came from the understanding that most patients with breast cancer have disseminated or “micrometastasized” tumors already at the time of diagnosis. Therefore, in order to efficiently fight the disease, a local surgical operation should be combined with effective simultaneous systemic treatment, such as radio-, hormonal or chemotherapy. While significant advances have been made with this so called adjuvant therapy, optimal therapy has not yet been defined for any breast cancer patients. One of the hurdles in the adjuvant therapy is that the tumor cells are either inherently resistant or develop resistance to such therapies. The underlying biochemical and genetic reasons of drug resistance in metastatic breast cancer are not clear. Hence, many women are given such adjuvant therapy, but only a minority will benefit.

Most therapy drugs are thought to work so that they activate self-destructive mechanisms in cancer cells and these cells therefore “commit suicide” (apoptosis) in response to the therapy. It has been hypothesized that resistant cancer cells somehow refuse to commit suicide in response to therapeutic drugs. The microarray technology could be used to study the genetic context of treatment resistance in breast cancer to improve the choice of an adequate therapy and our understanding of this biological phenomenon.

1.1.1.1 Tamoxifen Resistance Project

This project concerns the prediction of early distant metastases on Tamoxifen in early-stage breast cancer. The majority of early-stage breast cancers express estrogen receptors (ER) and receive Tamoxifen in the adjuvant setting. Yet up to 40% of these patients will relapse on Tamoxifen and develop incurable metastatic disease. Recent evidence from three large randomized controlled trials [Howell and Cuzick, 2005; Coombes and Hall, 2004; Goss and Ingle, 2003] exploring the role of aromatase inhibitor (AI) in the adjuvant setting shows a benefit from the novel strategy. However the optimal sequence and duration of Tamoxifen/AI treatment is unknown. Therefore, it is vital to learn to identify those women at higher risk of Tamoxifen resistance. The aim of this project is to identify genes that could predict for this subset of women.

In this thesis, we will focus on the analysis of gene expression profiles which are determined from 99 Tamoxifen-only treated ER positive early stage BC using Affymetrix© hgu133a and hgu133b chips (see Chapter 3). Within this group 30 (29%) patients developed distant recurrence at a median time[3] to relapse of 3.8 years and 75 (71%) remained disease free at a median of 10.7 years of follow-up. The independent validation set consisted of 69 ER+

[3] The median time is computed using the KM estimator (see Section 2.3.1).

Tamoxifen only treated breast cancer patients from a different institution (Karolinska, Sweden). Another independent dataset (Guys hospital, UK) consisted of 87 ER+ Tamoxifen only treated breast cancer patients.

Using these data, a group of genes will be selected to identify breast cancer patients at risk of early distant relapse on Tamoxifen. These patients could be the ideal candidates for upfront AIs, while the others would be considered for sequential Tamoxifen/AI.

1.2 Contributions

This section describes all the contributions presented in this thesis.

Methodology We propose a machine learning methodology based on machine learning methods (e.g. feature selection) and well-established survival statistics (e.g. statistical tests for the difference in survival between two groups). This methodology is sketched in figure 1.2 and includes methods for data preprocessing, feature selection, classifier construction and performance assessment (these methods will be described in Chapter 4).

Preprocessing Methods We introduce three new concepts in the microarray data preprocessing :

- Use of a normalization procedure (RMA [Irizarry et al., 2003a]) separately for each population of patients in order to facilitate the further analysis (inclusion of new populations during the analysis and an easier way to test new samples). See Section 4.2.2.2.
- A new correction method, called *population correction*, in order to minimize the variability due to the population effect. See Section 4.2.2.4.
- A prefiltering based on detection calls in order to discard noninformative probesets without using demographic data. Even if some measurements (MM probe intensities) are not taken into account by the normalization procedure (RMA), this information is used in the prefiltering based on detection calls. See Section 4.2.3.

Feature Selection We introduce a new feature selection method based on variable ranking, semi-supervised hierarchical clustering and cross-validation. See Section 4.3.

Classifier Validation on different Microarray Platforms We propose a new method to facilitate the classifier validation on different microarray patforms . This method is based on a specific feature construction. See Section 4.3.2.1.

Time-Dependent ROC Curve We use the recently introduced time-dependent ROC curves in breast cancer microarray studies in order to assess the classifier performance. Moreover, we provide an implementation of this method based on the R statistical tool [R Development Core Team, 2005]. See Section 4.5.4.

Cutoff Selection We introduce a new simple method to select a cutoff, based on the hazard ratio, for the risk scores. The aim of this method is to classify specifically a low-risk group including the smallest number of events before three years (early distant metastases). See Section 4.4.2.

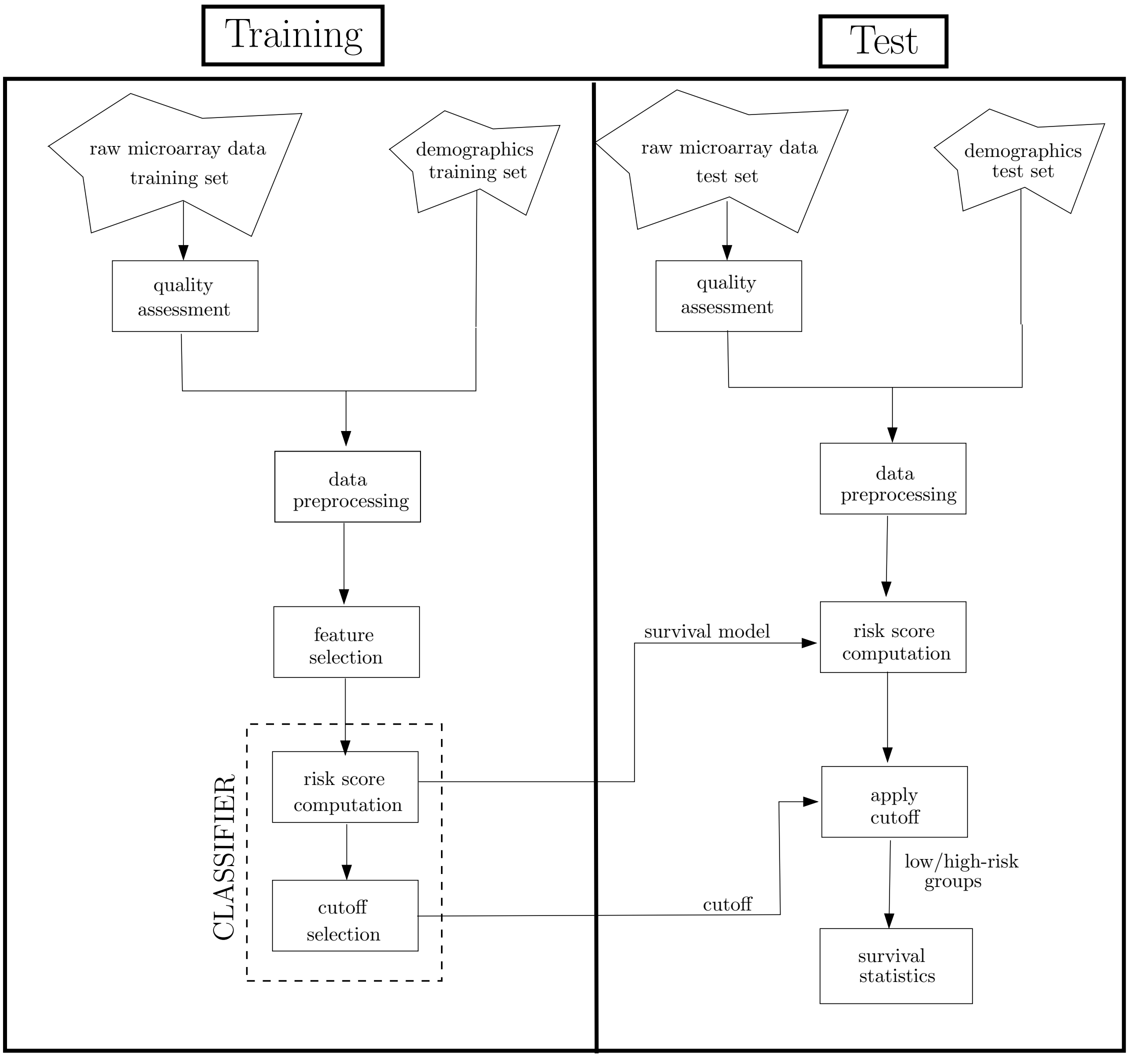


Figure 1.2: Machine learning methodology for survival analysis of microarray data.

1.3 Glossary

Adjuvant therapy Treatment given after the primary treatment to increase the chances of a cure. Adjuvant therapy may include chemotherapy, radiation therapy, hormone therapy, or biological therapy.

cDNA complementary DNA (cDNA) is single-stranded DNA synthesized from a mature mRNA template.

Consistency The consistency of an estimator means that it converges in probability to the true values as the sample gets larger, implying that the estimator is unbiased in large samples.

Covariate A covariate is a variable that is possibly predictive of the outcome under study. A covariate may be of direct interest or be a confounding variable or effect modifier.

Cross-hybridization The hydrogen bonding of a single-stranded DNA sequence that is partially but not entirely complementary to a single-stranded substrate. Often, this involves hybridizing a DNA probe for a specific DNA sequence to the homologous sequences of different species.

Cross-validation The cross-validation is the practice of partitioning a sample of data into subsets such that analysis is initially performed on a single subset, while further subsets are retained "blind" in order for subsequent use in confirming and validating the initial analysis.

Dendrogram A hierarchy representation by a dichotomous diagram, in which the end of a branch corresponds to an element and the level of a junction corresponds to the taxonomic distance from the two elements or the two groups that it connects.

Distant metastasis Cancer cells may spread to lymph nodes (regional lymph nodes) near the primary tumor. This is called nodal involvement, positive nodes, or regional disease. Cancer cells may spread to other parts of the body, distant from the primary tumor. If a new cancer grows in such sites, we call it a distant metastasis.

Expressed Sequence Tag A short strand of DNA that is a part of a cDNA molecule and can act as identifier of a gene.

GenBank The GenBank sequence database is an annotated collection of all publicly available nucleotide sequences and their protein translations. This database is produced at National Center for Biotechnology Information (NCBI) as part of an international collaboration with the European Molecular Biology Laboratory (EMBL) Data Library from the European Bioinformatics Institute (EBI) and the DNA Data Bank of Japan (DDBJ).

Gene Expression Transcription of the information contained within the DNA into messenger RNA (mRNA) molecules that are then translated into proteins.

Hybridization Hybridization is the process of binding complementary pairs of DNA molecules. A DNA molecule has a very strong preference for its sequence complement, so just mixing complementary sequences is enough to induce them to hybridize. Hybridization is temperature dependent, so DNAs that hybridize strongly at low temperature can be temporarily separated (denatured) by heating.

Location parameter The location parameter simply shifts the distribution left or right on the horizontal axis.

Longitudinal data Observations collected over a period of time.

Lymph nodes Lymph nodes are components of the lymphatic system. Clusters of lymph nodes are found in the underarms, groin, neck, chest, and abdomen. Lymph nodes act as filters, with an internal honeycomb of connective tissue filled with lymphocytes that collect and destroy bacteria and viruses. When the body is fighting an infection, these lymphocytes multiply rapidly and produce a characteristic swelling of the lymph nodes.

Mer Monomeric Unit. The largest constitutional unit, coming from only one molecule of a monomer in a process of polymerization.

Meta-analysis Analysis involving several sources of microarray data (e.g. AFFYMETRIX$^{©}$ and AGILENT$^{©}$ data).

Microarray Ordered arrangement, on a miniaturized support of glass, of silicon or polymer, of hundreds or thousands of molecular probes whose nucleotidic sequence is known, and whose function is to recognize, in a mixture, their complementary nucleotidic sequences.

Monotone function The function f is monotone if, whenever $x \leq y$, then $f(x) \leq f(y)$. Stated differently, a monotone function is one that preserves the order.

Neo-adjuvant therapy Treatment also known as primary systemic therapy, or primary medical therapy: when chemotherapy is given before primary surgery.

Oligonucleotide Short fragment of a single-stranded DNA.

Prognosis Prediction of survival independently of treatment.

Polymerase Chain Reaction (PCR) Exponential amplification of almost any region of a selected DNA molecule.

Probe Easily detectable molecule which has the property to be located specifically either on another molecule, or in a given cellular compartment. Various molecules can be used as probe with condition that a marker (enzyme, compound radioactive or fluorescent) can be associated with the probe which allows its detection. Generally the probe is a nucleic acid fragment (ARN or ADN).

Probeset Set of probes used in the microarray platform of AFFYMETRIX$^{©}$. Even if, generally, a probeset corresponds to one gene, the expression of one gene may be measured by a set of probesets.

Reverse Transcriptase Polymerase Chain Reaction (RT-PCR) Molecular technique which uses upon the reverse transcriptase to amplify a sequence of RNA and to transform it into DNA.

Scale parameter The effect of a scale parameter greater than one is to stretch the PDF. The greater the magnitude, the greater the stretching. The effect of a scale parameter less than one is to compress the PDF. The compressing approaches a spike as the scale parameter goes to zero. A scale parameter of 1 leaves the PDF unchanged (if the scale parameter is 1 to begin with) and non-positive scale parameters are not allowed.

Sensitivity The sensitivity of a binary classification test is a parameter that expresses something about the test's performance. The sensitivity of such a test is the proportion of those cases having a positive test result of all positive cases tested ($\frac{TP}{TP+FN}$).

Shape parameter Many probability distributions are not a single distribution, but are in fact a family of distributions. This is due to the distribution having one or more shape parameters. Shape parameters allow a distribution to take on a variety of shapes, depending on the value of the shape parameter. These distributions are particularly useful in modeling applications since they are flexible enough to model a variety of datasets.

Skewness Skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable. Roughly speaking, a distribution has positive skew (right-skewed) if the higher tail is longer and negative skew (left-skewed) if the lower tail is longer.

Specificity The specificity of a binary classification test is a parameter that expresses something about the test's performance. The specificity of such a test is the proportion of true negatives of all the negative samples tested ($\frac{TN}{TN+FP}$).

Tamoxifen A drug (NOLVADEX©) used to treat breast cancer, and to prevent it in women who are at a high risk of developing breast cancer. TAMOXIFEN blocks the effects of the hormone estrogen in the breast. It belongs to the family of drugs called antiestrogens.

1.4 Abbreviations and Acronyms

AUC Area Under the Curve.

AI Aromatase Inhibitor.

AFT Accelerated Failure Time.

BIG Breast International Group.

CEL CELl intensities.

CDF Cumulative Distribution Function.

DCIS Ductal Carcinoma In Situ.

DDBJ DNA Data Bank of Japan.

DF Degree of Freedom.

DMFS Distant Metastases Free Survival.

EBI European Bioinformatics Institute.

EMBL European Molecular Biology Laboratory.

EORTC-BCG Breast Cancer Group of the European Organization for Research and Treatment in Cancer.

ER Estrogen Receptor.

EST Expressed Sequence Tag.

FN False Negatives.

FP False Positives.

GCOS GeneChip Operating Software.

GO Gene Ontology.

GUYT Population of Tamoxifen treated patients coming from the Guys hospital.

HR Hazard Ratio.

IJB Institut Jules Bordet.

KIT Population of Tamoxifen treated patients coming from the Karolinska hospital.

KM Kaplan-Meier.

KNN K-Nearest Neighbours.

LASSO Least Absolute Shrinkage and Selection Operator.

LN Lymph Node.

LOO Leave-One-Out.

MAS Microarray Affymetrix© Suite.

MGED Microarray Gene Expression Data Society.

MIAME Minimum Information About a Microarray Experiment.

MM Mis-Match.

NCBI National Center for Biotechnology Information.

OXFT Population of Tamoxifen treated patients coming from the John Radcliffe hospital hospital.

PDF Probability Density Function.

PM Perfect Match.

RMA Robust Multi-array Average.

ROC Receiving Operator Characteristic.

RT-PCR Reverse Transcriptase Polymerase Chain Reaction.

SIB Swiss Institute of Bioinformatics.

SVM Support Vector Machines.

TN True Negatives.

TP True Positives.

1.5 Notations

N Number of samples.

P Number of probes.

Q Number of probesets.

F Number of features.

n Number of input variables.

Pop Set of populations.

X Set of covariates.

G Indicator variable for group ($G = 0$ for the low-risk group and $G = 1$ for the high-risk group).

$\mathcal{S}$ Scoring function.

$X, Y, \ldots$ Upper case letters represent random variables (except the previously defined "number of ...").

$x, y, \ldots$ Lower case letters represent the realization of random variables (except n).

$\mathbf{x}, \mathbf{X}, \boldsymbol{\beta}, \ldots$ Bold letters represents vectors or matrices.

T_i Time of occurrence/censoring for the sample i ($i \in \{1, \ldots, N\}$).

δ_i Indicator status for sample i ($i \in \{1, \ldots, N\}$).

β Coefficients of a linear regression model.

$\hat{\beta}$ Estimated coefficients.

D_N Dataset of N samples $\{x_i, y_i\}$ ($i \in \{1, \ldots, N\}$).

Chapter 2

Survival Analysis

Contents

Survival Analysis is a class of statistical methods for studying the occurrence and timing of events. These methods are most often applied to the study of deaths but can treat different kinds of event including the onset of disease, equipment failures, arrests, etc.

Survival analysis was designed for longitudinal data on the occurrence of events. An event can be defined as a qualitative change[1] that can be situated in time. For instance a disease consists of a transition from an healthy state to a diseased state.

Moreover, the timing of the event is also considered for analysis. Ideally, the transitions occur virtually instantaneously and the exact times at which the event occurs is known. Some

[1] A qualitative change is defined as a transition from one discrete state to another.

transitions may take a little time, however, and the exact time of onset may be unknown or ambiguous.

For survival analysis, the best observation plan is *prospective*. By prospective we mean that the observation of a set of individuals starts at some well-defined point in time and they are followed for some substantial period of time, recording the time at which the events of interest occur.

In this thesis, survival analysis is used with *retrospective* data, looking back at patients' medical history. These data present some potential limitations :

- the data are prone to errors, some events may be forgotten
- the sample of patients may be a biased subsample of the initial population of interest.

Survival data have two common features that are difficult to handle with conventional statistical methods : *censoring* and *time-dependent covariates* (sometimes called time-varying explanatory variables). Consider the following example, which illustrates both these problems. A sample of 432 inmates released in Maryland state prisons was followed for one year after release [Rossi et al., 1980]. The event of interest was the first arrest. The aim was to determine how the occurrence and timing of arrests depended on several covariates (predictor variables). Some of these covariates (like age of release and number of previous convictions) remained constant over the one-year interval. Others (like marital status and employment status) could change at any time during the follow-up period.

If we narrow our focus on a dichotomous dependent variable (arrested or not arrested), conventional methods that could analyze such data, are the logistic regression (logit) [McCullagh and Nelder, 1989], linear discriminant analysis or support vector machines for instance (see [Duda et al., 2001] for a review of such classification methods). But this analysis ignores information on the timing of arrest. It is natural to suppose that people who are arrested one week after release have, on average, a higher propensity to be arrested than those who are not arrested until the 52nd week. At least, ignoring that information should reduce the precision of the estimates.

One solution to this problem is to make the length of time between release and first arrest the dependent variable and then estimate it by a conventional linear regression [McCullagh and Nelder, 1989]. But it remains a problem with persons who were not arrested during the one-year follow-up. Such cases are referred to as *censored*. A couple of obvious ad-hoc methods exist for dealing with censored cases, but neither works well. One method is to discard the censored cases but this proportion may be large. This method may result in large biases. Alternatively, the time of arrest could be set at one year for all those who were not arrested. That is clearly an underestimate, however, and some of those ex-convicts may never be arrested. Again large biases may occur.

Whichever method is used, it is not clear how a time-dependent variable like employment status can be appropriately incorporated into either the classification methods for the occurrence of arrests or the linear model for the timing of arrests.

The methods of survival analysis allow for censoring and many also allow for time-dependent covariates in combining the information with the censored and the uncensored cases [Allison, 1995].

2.1 Censoring Data

An observation on a random variable T is right-censored if all you know about T is that it is greater than some value c. In survival analysis, T is typically the time of occurrence for some event, and cases are right-censored because observation is terminated before the event occurs.

The simplest and the most common situation is depicted in Figure 2.1. Suppose that this figure reports some of the data from a study in which all persons receive heart surgery at time 0 and are followed for 3 years thereafter. The horizontal axis represents time. Each of the horizontal lines labeled A through E represents a single person. An X indicates that a death occurred at that point in time. The vertical line at 3 is the point at which the follow-up of the patients is stopped. Any death occurring at time 3 or earlier are observed and, hence, those death times are uncensored. Any deaths occurring after 3 years are not observed, and those death times are censored at time 3.

Therefore, persons A, C and D have uncensored death times, while person B and E have right-censored death times. Observations that are censored in this way are referred to as *singly right-censored.* Singly refers to the fact that all the observations had the same censoring time. Observations that are not censored are said to have a censoring time, in this case three years. It is just that their death times did not exceed their censoring time. Moreover, the censoring time is fixed and is under the control of the investigator.

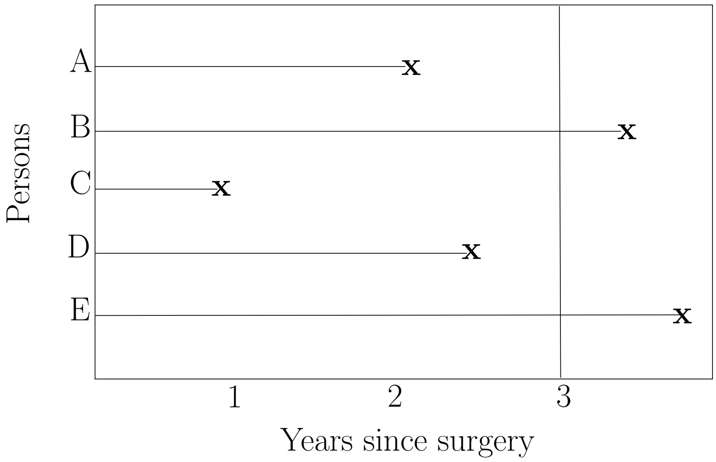


Figure 2.1: Singly right-censored data.

Random censoring occurs when observations are terminated for reasons that are not under the control of the investigator. This situation can be illustrated by the following example : in a study of divorces, a sample of couples are followed for 10 years beginning with the marriage and the timing of all divorces are recorded. Clearly, couples that are still married after ten years are censored by a mechanism identical to this applied for the singly right-censored data. But for some couples, either the husband or the wife may die before the ten years are up. Some couples may move out and it may be impossible to contact them. Still other couples may refuse to participate after, say five years. These kinds of censoring are depicted in Figure 2.2 where the O for the couples B and C indicates that observation is censored at that point in time.

Random censoring can also be produced when there is a single termination time, but

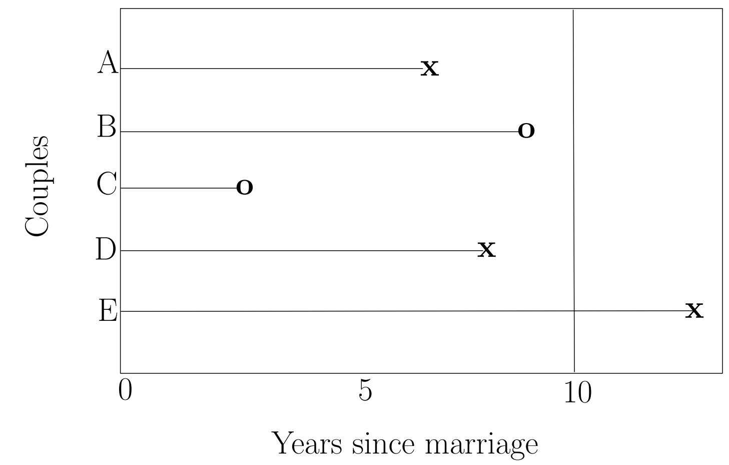


Figure 2.2: Randomly censored data.

entry times vary randomly across individuals. Consider again the example in which people are followed for heart surgery until death. A more likely scenario is one in which people receive heart surgery at various point in time, but the study has to be terminated on a single date. All persons still alive on that date are considered censored, but their survival time from surgery will vary. This censoring is considered random because the entry times are typically not under the control of the investigator.

Standard methods of survival analysis treat the right-censored data but require that random censoring be *noninformative*. Here is how this situation is described in [Cox and Oakes, 1984] :

> A crucial condition is that, conditionally on the values of any explanatory variables, the prognosis for any individual who has survived to c_i should not be affected if the individual is censored at c_i. That is, an individual who is censored at c should be representative of all those subjects with the same values of he explanatory variables who survive to c (p. 5).

The best way to understand this condition is to think about possible violations. In the divorce example mentioned earlier, it is plausible that those couples who refuse to continue participating in the study are more likely to be experiencing marital difficulties and, hence, are at greater risk of divorce. The censoring is informative assuming that measured covariates do not fully account for the association between drop-out and marital difficulty. Informative censoring can, at least in principle, lead to severe biases, but it is difficult in most situations to assess the magnitude or direction of those biases.

In this thesis we will focus on analysis of right-censored data.

2.2 Survival Distributions

The standard approaches to survival analysis are based on statistical modeling. The times at which events occur are assumed to be realizations of some random variable T. Three ways of describing the probability distribution of T are presented in this section :

1. the cumulative distribution function

2. the probability density function
3. the hazard function.

2.2.1 Cumulative Distribution Function

The cumulative distribution function (CDF) of a random variable T, denoted by $F(t)$, is a function giving the probability that the variable will be less than or equal to any specific value t, i.e. $F(t) = \Pr\{T \leq t\}$. In survival analysis, it is more common to work with the *survivor function*, defined as $S(t) = \Pr\{T > t\} = 1 - F(t)$. If the event of interest is a death, the survivor function gives the probability of surviving beyond t. Because T cannot be negative, $S(0) = 1$.

2.2.2 Probability Density Function

When variables are continuous, another useful way of describing the probability distribution is the probability density function (PDF). This function is defined as

$$f(t) = \frac{dF(t)}{dt} = -\frac{dS(t)}{dt} \tag{2.1}$$

2.2.3 Hazard Function

In the case of continuous survival data, the *hazard function* is actually more used than the PDF in order to describe distributions. The hazard function is defined as

$$h(t) = \lim_{\Delta t \to 0} \frac{\Pr\{t \leq T < t + \Delta t \,|\, T \geq t\}}{\Delta t} \tag{2.2}$$

The function $h(t)$ quantifies the instantaneous risk[2] that an event will occur in the small interval between t and $t + \Delta t$. The probability in the numerator of (2.2) is conditional on the individual surviving to time t because individuals who have already experienced the event should not be considered.

The definition of the hazard function in (2.2) is similar to an alternative definition of the PDF

$$f(t) = \lim_{\Delta t \to 0} \frac{\Pr\{t \leq T < t + \Delta t\}}{\Delta t} \tag{2.3}$$

The only difference is that the probability in the numerator of (2.3) is an unconditional probability, whereas the probability in (2.2) is conditional on $T \geq t$. For this reason, the hazard function is sometimes described as a *conditional density*.

The survivor function, the probability density function and the hazard function are equivalent ways of describing a continuous probability distribution. The relationship between the PDF and the survivor function is given directly by the (2.1). Another simple formula expresses the hazard function in terms of the PDF and the survivor function :

$$h(t) = \frac{f(t)}{S(t)} \tag{2.4}$$

[2] Although it may be helpful to think of the hazard as the instantaneous probability of an event at time t, this quantity is not a probability and may be greater than 1. This is due to the division by Δt in (2.2). Although the hazard has no upper bound, it cannot be less than 0.

Together, (2.4) and (2.1) imply that

$$h(t) = -\frac{d}{dt} \log S(t) \tag{2.5}$$

By integrating both sides of (2.5), we obtain an expression of the survivor function in terms of the hazard function :

$$S(t) = \exp\left\{-\int_0^t h(u)du\right\} \tag{2.6}$$

Together with (2.4), this formula leads to

$$f(t) = h(t) \exp\left\{-\int_0^t h(u)du\right\} \tag{2.7}$$

The hazard is a dimensional quantity that has the form *number of events per interval of time*. This is why the hazard is sometimes called a *rate*. The units in which time is measured must be known in order to interpret the value of the hazard. Suppose that the hazard of contracting influenza at some particular point in time is 0.015 with time measured in months. This means that if the hazard stays at that value during a period of one month, one expects that a person will contract the influence 0.015 times during that month.

2.2.4 Simple Hazard Models

The hazard function is a useful way of describing the probability distribution for the time of event occurrence. Every hazard function has a corresponding probability distribution. This section examines some rather simple hazard functions and discusses their associated probability distributions.

The simplest hazard functions specifies that the hazard is constant over time, that is, $h(t) = \lambda$ or, equivalently $\log h(t) = \mu$. Substituting this hazard into (2.6) and carrying out the integration implies that the survival function is $S(t) = e^{-\lambda t}$. From (2.1), we get the PDF $f(t) = \lambda e^{-\lambda t}$. This is the PDF for the exponential distribution with parameter λ. Thus, a constant hazard implies an exponential distribution for the time until an event occurs (or the time between events).

Let now the natural logarithm of the hazard be a linear function of time :

$$\ln h(t) = \mu + \alpha t$$

where μ and α are real constant values. Taking the logarithm is a convenient way to ensure that $h(t)$ is nonnegative, regardless of the value of μ, α and t. We can rewrite the equation as

$$h(t) = \lambda \gamma^t$$

where $\lambda = e^\mu$ and $\gamma = e^\alpha$. This hazard function implies that the time of event occurrence has a *Gompertz* distribution (see Figures 2.3, 2.4 and the Table 2.2 for the Gompertz distribution, the Gompertz hazard function and the Gompertz PDF formula respectively). Alternatively we can assume that

$$\ln h(t) = \mu + \alpha \ln t$$

which can be rewritten as

$$h(t) = \lambda t^\alpha$$

with $\lambda = e^{\mu}$. This equation implies that the time of event occurrence follows a *Weibull* distribution (see figures 2.5, 2.6 and the Table 2.2 for the Weibull distribution, the Weibull hazard function and the Weibull PDF formula respectively).

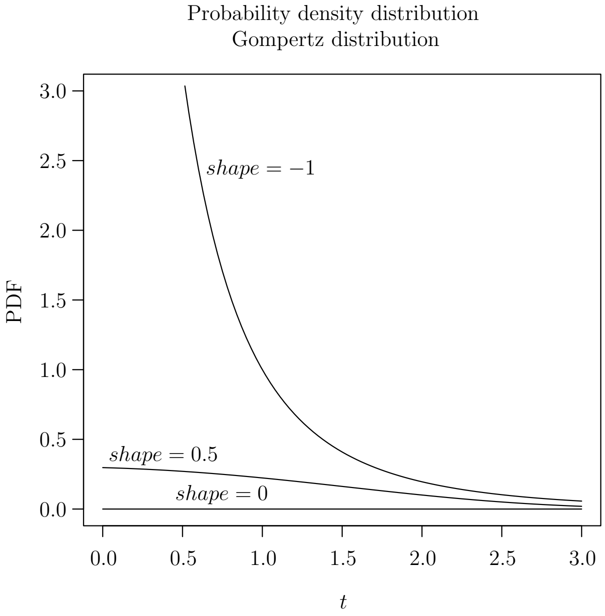


Figure 2.3: Gompertz distribution for time of event occurrence. The probability density distribution is given for different values of the *shape* parameter (the *shape* corresponds to the α parameter of the Gompertz hazard function, such that $shape = \alpha$).

The *Gompertz* and the *Weibull* distributions coincide with the exponential distribution in the special case $\alpha = 0$. When α is not zero, the hazard is either always decreasing or always increasing with time for both distributions. One difference between them is that, for the Weibull model, when $t = 0$, the hazard is either zero or infinite. With the Gompertz model, the initial value of the hazard is λ, which can be any nonnegative number.

We can extent each of these models to allow for the influence of covariates. For instance, a covariate for the situation reported by the Figure 2.1 could be the age of the patient at time of surgery or its blood group. Thus, if we have covariates $x_1, x_2, \ldots, x_k$, we can write

$$\begin{aligned} Exponential &: \quad \ln h(t) = \mu + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k && (2.8) \\ Gompertz &: \quad \ln h(t) = \mu + \alpha t + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k && (2.9) \\ Weibull &: \quad \ln h(t) = \mu + \alpha \ln t + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k && (2.10) \end{aligned}$$

2.3 Estimating Survival Curves

Prior to 1970, the estimation of $S(t)$ was the predominant method of survival analysis [Gross and Clark, 1975]. Nowadays, the workhorse of the survival analysis is the Cox regression method [Cox, 1972]. Nevertheless, survival curves are still useful for preliminary examination of the data, for computing derived quantities from regression models (e.g. the median survival time or the five-year probability of survival) and for evaluating the fit of regression models.

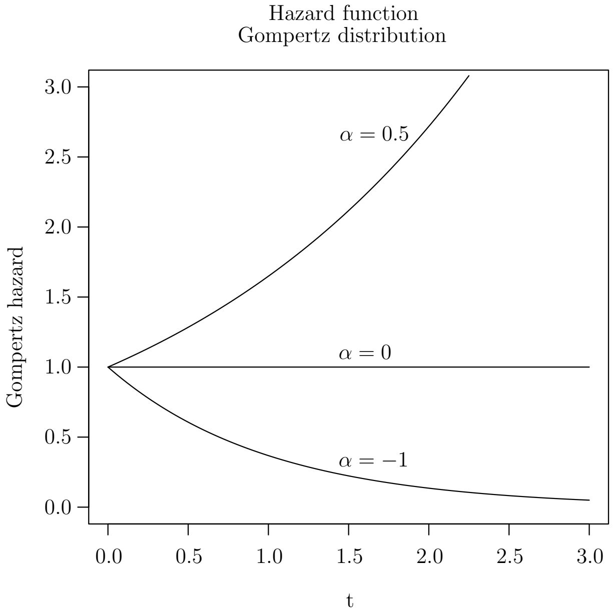


Figure 2.4: Typical hazard functions ($h(t) = \lambda\gamma^t$ with $\lambda = 1$) for the Gompertz distribution.

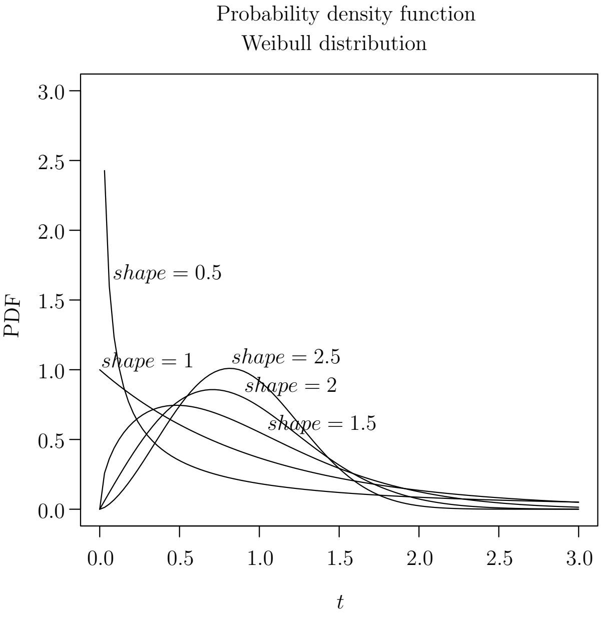


Figure 2.5: Weibull distribution for time of event occurrence. The probability density distribution is given for different values of the *shape* parameter (the *shape* corresponds to the α parameter of the Weibull hazard function, such that $shape = \alpha + 1$).

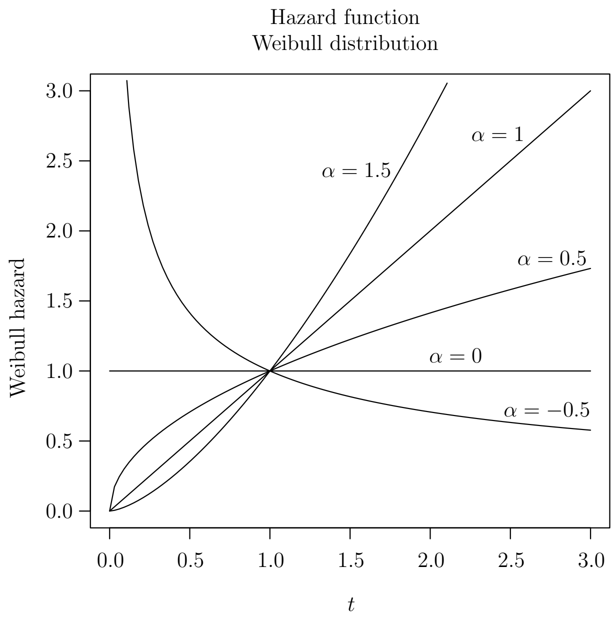


Figure 2.6: Typical hazard functions ($h(t) = \lambda t^{\alpha}$ with $\lambda = 1$) for the Weibull distribution.

There exist two methods to estimate survivor functions : the *Kaplan-Meier* and the *life-table* methods. The *Kaplan-Meier* method is most suitable for small datasets with precisely measured event times. The *life-table* method may be better for large datasets or when the measurement of event times is crude [Allison, 1995].

In this thesis, the number of samples is small (high feature/sample ratio that will be described in Section 2.6). It is the reason why the life-table method will not be treated.

2.3.1 Kaplan-Meier Method

The *Kaplan-Meier* (KM) estimator is the most widely used method for estimating survivor functions. Also known as the *product-limit estimator*, Kaplan and Meier have shown in 1958 that this estimator is the nonparametric maximum likelihood estimator [Kaplan and Meier, 1958].

When there are no censored data, the KM estimator is simple and intuitive.We have seen in Section 2.2 that the survivor function $S(t)$ is the probability that an event time is greater than t, where t can be any nonnegative number. In the case of no censoring, the KM estimator is just the sample proportion of observations with event time greater than t.

If data are right censored, the observed proportion of cases with event times greater than t can be biased downward because cases that are censored before t may have experienced an event before t without our knowledge. Suppose there are r distinct event times, $t_1 < t_2 < \cdots < t_r$. At each time t_j, there are n_j individuals who are said to be at risk of an event. *At risk* means they have not experienced an event nor have they been censored prior to time t_j. If any cases are censored at exactly t_j, they are also considered to be at risk at t_j. Let d_j be the number of individuals who die at time t_j. The KM estimator is then defined as

$$\widehat{S}(t) = \prod_{j:t_j \leq t} \left[1 - \frac{d_j}{n_j}\right] \tag{2.11}$$

for $t_1 \leq t \leq t_r$. In words, the quantity in brackets can be interpreted as the conditional probabilities of surviving to time t_{j+1}, given that one has survived to time t_j. So, $\widehat{S}(t)$ is the probability to survive to time t. For $t < t_1$ (the smallest event time), $\widehat{S}(t)$ is defined to be 1. For $t > t_r$ (the largest observed event time), the definition of $\widehat{S}(t)$ depends on the configuration of the censored observations. When there are no censored times greater than t_r, $\widehat{S}(t)$ is set to $\widehat{S}(t_r)$ for $t > t_r$. When there are censored times greater than t_r, $\widehat{S}(t)$ is undefined for t greater than the largest censoring time.

Here is a small example concerning the survival of breast cancer patients (inspired from [Collett, 2003]). Consider the data in Table 2.1.

Patient id	Survival time (in months)	Event
1	5	1
2	8	1
3	10	0
4	13	1
5	18	0

Table 2.1: Survival times for breast cancer patients.

The corresponding survival curve using KM estimator is given in Figure 2.7.

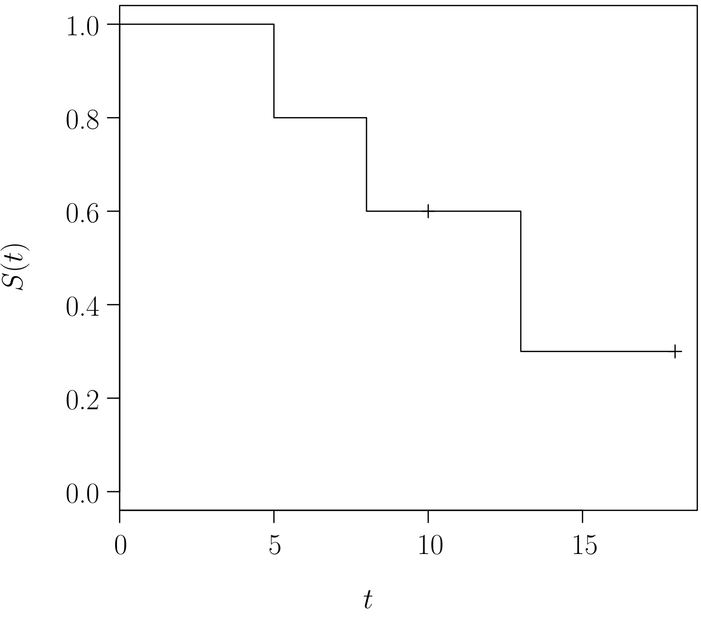


Figure 2.7: Survival curve estimated by the KM estimator from data in Table 2.1. The "+" represents the censoring.

An estimate of standard error of the KM estimate can be obtained by the Greenwood's formula [Greenwood, 1926; Collett, 2003] :

$$\begin{aligned}
\hat{\sigma}_G^2\left\{\widehat{S}(t)\right\} &= \{\widehat{S}(t)\}^2 \sum_{j:t_j \leq t} \frac{d_j}{n_j(n_j - d_j)} \\
\widehat{\mathrm{se}}_G\left\{\widehat{S}(t)\right\} &= \widehat{S}(t)\sqrt{\sum_{j:t_j \leq t} \frac{d_j}{n_j(n_j - d_j)}} \qquad (2.12)
\end{aligned}$$

This is derived by estimating each term in the product expansion of $\widehat{S}(t)$ separately. Moreover, the bootstrap method can be used to estimate the variance of $\widehat{S}(t)$ [Akritas, 1986]. It can be shown that the KM estimator is asymptotically normal according the sample size, with mean $\widehat{S}(t)$ and variance estimated by the Greenwood's formula [Meier, 1975]. Intervals of confidence around KM estimates can be computed using these results.

2.4 Estimating Regression Models

Survivor functions can be estimated by regression models. In survival analysis, there exist two categories of such regression models : the *parametric* and the *semiparametric* regression models.

2.4.1 Parametric Regression Models

The parametric regression models with censored data are estimated using the method of *maximum likelihood.* Such class of regression models is known as the *accelerated failure time* (AFT) class. In the most general form, the AFT model describes a relationship between the survivor functions of any two individuals. If $S_i(t)$ is the survivor function for individual i, then for any other individual j, the AFT model holds that

$$S_i(t) = S_j(\phi_{ij}t)$$

where $i, j \in \{1, \ldots, N\}$ and ϕ_{ij} is a constant that is specific to the pairs (i, j). This model says that what makes different an individual from another is the rate at which they age. A good example is the conventional wisdom that a year for a dog is equivalent to seven years for a human.

In practice, the models commonly used are a special case of the AFT model that is quite similar in form to an ordinary linear regression model. Let T_i be a random variable denoting the event time for the ith individual in the sample, and let $x_{i1}, x_{i2}, \ldots, x_{in}$ be the values of n covariates for that same individual. The model is then

$$\ln T_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_n x_{in} + \epsilon_i \qquad (2.13)$$

where ϵ_i is a random disturbance term, and $\beta_0, \beta_1, \ldots, \beta_n$ are parameters to be estimated.

In a linear regression model, it is typical to assume that ϵ_i has a normal distribution with a mean and variance that are constant over i, and that the ϵ's are independent across observations. It is the case for one member of the AFT class, the *log-normal model*[3]. However there exist several alternatives allowing distributions of ϵ besides the normal distribution but

Distribution of ϵ	Distribution of T	PDF of T
extreme value (2 par.)	Weibull	$\frac{a}{b}\left(\frac{x-c}{b}\right)^{a-1} e^{-\left(\frac{x-c}{b}\right)^{a}}$ $x \geq c;\ a, b > 0$
extreme value (1 par.)	exponential	$\frac{1}{b}e^{\frac{-(x-c)}{b}}$ $x \geq c;\ b > 0$
log-gamma	gamma	$\frac{\left(\frac{x-c}{b}\right)^{a-1} e^{\left(-\frac{x-c}{b}\right)}}{b\Gamma(a)}$ $x \geq c;\ a, b > 0$
logistic	log-logistic	$\frac{a}{b}\frac{\left(\frac{x-c}{b}\right)^{a-1}}{\left[1+\left(\frac{x-c}{b}\right)^{a}\right]^{2}}$ $x \geq c;\ a, b > 0$
normal	log-normal	$\frac{e^{-\left(\left(\ln\left(\frac{(x-c)}{m}\right)\right)^{2}/(2a^{2})\right)}}{(x-c)a\sqrt{2\pi}}$ $x \geq c;\ a, b > 0$

Table 2.2: Alternatives for the distributions of ϵ and their corresponding distributions of T. Legend : a is the shape parameter, b is the scale parameter and c is the location parameter.

retaining the assumptions of constant mean and variance, as well as independence across observations. Some example of these alternatives are given in Table 2.2.

The main reason of the use of such alternatives is that they have different implications for the hazard functions that may lead to different substantive interpretations.

The parameters of such models are estimated by the maximum likelihood method (see Section 2.4.2).

Recently, the parametric regression models have been eclipsed by the semiparametric regression model with the famous Cox regression model. this is why this thesis will focus on this promising method.

2.4.2 Semiparametric Regression Models

The semiparametric regression models refer to the method first proposed in 1972 by the British statistician Sir David Cox in his famous paper “Regression Models and Life Tables” [Cox, 1972]. It is difficult to exaggerate the impact of this paper. In the 1992 *Science Citation Index*, it was cited over 800 times, making it the most highly cited journal article in the entire literature of statistics. In fact, [Garfield, 1990] reported that its cumulative citation count placed it among the top 100 papers in all of science.

This enormous popularity can be explained by the fact that, unlike the parametric methods, Cox's method does not require the selection of some particular probability distribution to represent survival times. For this reason, the method is called *semiparametric*. Moreover, this method makes it relatively easy to incorporate time-dependent covariates[4].

2.4.2.1 The Proportional Hazards Model

In his 1972 paper, Cox made two significant innovations. First, he proposed a model that is standardly referred as the *proportional hazards model*[5]. Second, he proposed a new estimation

[3] This model is called the log-normal model because if $\ln T$ has a normal distribution, then T has a log-normal distribution.

[4] The time-dependent covariates are covariates which value may change over the course of the observation period.

[5] It is important to mention that the model proposed by Cox can be generalized to allow for nonproportional hazards.

method that was later named *maximum partial likelihood*. The term *Cox regression* refers to the combination of the model and the estimation method.

Model Let's start with the basic model that does not include time-dependent covariates or nonproportional hazards. The model is usually written as

$$h_i(t) = \lambda_0(t) \exp\left(\beta_1 x_{1i} + \cdots + \beta_n x_{ni}\right) \tag{2.14}$$

This equation says that the hazard for individual i at time t is the product of two factors :

- a baseline hazard function $\lambda_0(t)$ that is left unspecified, except that it can not be negative
- a linear function of a set of n fixed covariates, which is exponentiated.

The function $\lambda_0(t)$ can be regarded as the hazard function for an individual whose covariates all have values of zero.

Taking the logarithm of both sides of (2.14), we can rewrite the model as

$$\ln h_i(t) = \alpha(t) + \beta_1 x_{1i} + \cdots + \beta_n x_{ni} \tag{2.15}$$

where $\alpha(t) = \ln \lambda_0(t)$. If we further specify $\alpha(t) = \alpha$, we get the exponential model with covariates (2.8). If we specify $\alpha(t) = \alpha t$, we get the Gompertz model. Finally, if we specify $\alpha(t) = \alpha \ln t$, we have the Weibull model (see Section 2.2.4). As we will see, however, the great attraction of Cox regression is that such choices are unnecessary. The function $\alpha(t)$ can take any form whatever.

This model is called the proportional hazards model because the hazard for any individual is a fixed proportion of the hazard for any other individual. It can be shown by taking the ratio of the hazards for two individuals i and j for $i, j \in \{1, \ldots, N\}$, and applying (2.14)

$$\frac{h_i(t)}{h_j(t)} = \exp\left\{\beta_1(x_{1i} - x_{1j}) + \cdots + \beta_n(x_{ni} - x_{nj})\right\} \tag{2.16}$$

We can see in (2.16) that $\lambda_0(t)$ cancels out of the numerator and denominator. As a result, the ratio of the hazards for any two individuals is constant over time. If we graph the ln hazards for any two individuals, the proportional hazards property implies that the hazard functions should be strictly parallel as depicted in Figure 2.8.

Estimation Fitting the proportional hazards model given in (2.14) to an observed set of survival data entails estimating the unknown coefficients, $\beta_1, \beta_2, \ldots, \beta_n$, of the covariates $X_1, X_2, \ldots, X_n$, in the linear component of the model. The baseline hazard function $\lambda_0(t)$ may also need to be estimated. It turns out that these two components of the model can be estimated separately. The β's are estimated first and these estimates are then used to construct an estimate of the baseline hazard function (see [Collett, 2003] for details about the estimation of the baseline hazard function). This is an important result, since it means that in order to make inferences about the effect of n covariates, $X_1, X_2, \ldots, X_n$, on the relative hazard, $h_i(t)/\lambda_0(t)$, we do not need an estimate of $\lambda_0(t)$.

Since the estimation of β's does not take into account the baseline hazard function, the resulting estimates are not fully efficient. This means that their standard errors are larger than they would be with the entire likelihood function. However, the loss of efficiency is quite

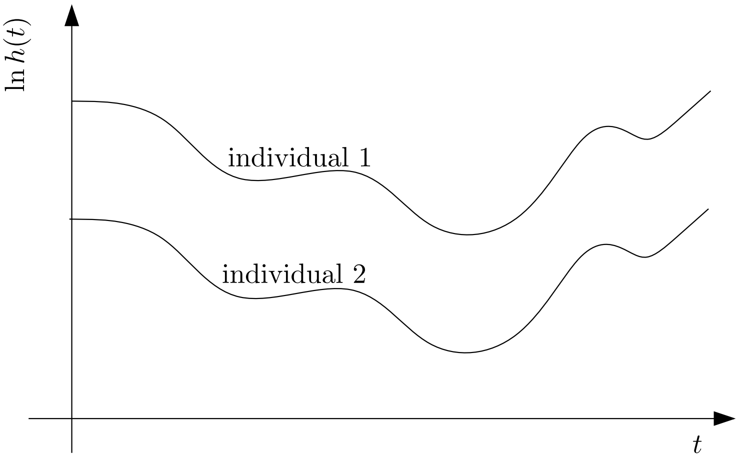


Figure 2.8: Parallel hazard functions from the proportional hazard model.

small in most cases [Efron, 1977]. In return, estimates have good properties regardless of the actual shape of the baseline hazard function. Partial likelihood estimates have still two of the three standard properties of maximum likelihood estimates : they are consistent and asymptotically normal[6] [Cox, 1972].

Another interesting property of partial likelihood estimates is that they depend only on the ranks of the event times, not their numerical values. This implies that any monotone transformation of the event times will leave the coefficient estimates unchanged.

Using the same notation as before, we have N independent individuals ($i \in \{1, \ldots, N\}$). For each individual i, the data consist on three parts : t_i, δ_i and $\mathbf{x}_i$, where t_i is the time of the event or the time of censoring, δ_i is an indicator variable with a value of 1 if t_i is uncensored or a value of 0 if t_i is censored, and $\mathbf{x}_i = [x_{1i}, x_{2i}, \ldots, x_{ni}]$ is a vector of n covariate values.

An ordinary likelihood function is typically written as a product of the likelihoods for all the individuals in the sample. On the other hand, the partial likelihood can be written as a product of the likelihoods for all the events that are observed. So we can write

$$PL = \prod_{i=1}^{N} L_i \tag{2.17}$$

where L_i is the likelihood for the ith event. Next we need to know how the individuals L_i are constructed. This is best explained by way of an example. Consider the data in Table 2.1 where we add a column to for a covariate X. The covariate X has a value of 1 if the tumor had a positive marker for distant metastasis, 0 otherwise (see Table 2.3).

The first event occurred to patient 1 in month 5. To construct the partial likelihood L_1 for this event, we ask the following question : “Given that an event occurred in month 5, what is the probability that it happened to patient 1 rather than any other patients ?”. The answer is the hazard for patient 1 at month 5 divided by the sum of the hazards for all the patients who were at risk in that same month :

$$L_1 = \frac{h_1(5)}{h_1(5) + h_2(5) + \cdots + h_5(5)} \tag{2.18}$$

[6] Partial likelihood estimates are approximately unbiased and their sampling distribution is approximately normal in large samples.

Patient id	Survival time (in months)	Event	X
1	5	1	1
2	8	1	1
3	10	0	1
4	13	1	0
5	18	0	0

Table 2.3: Survival times for breast cancer patients with the covariate X.

While this expression has considerable intuitive appeal, the derivation is actually rather involved and will not presented here (see [Collett, 2003] for details).

The second event occurred to patient 2 in month 8. Patient 1 is no longer at risk of event because he had already an event. So L_2 has the same form as L_1, but the hazard for patient 1 is removed from the denominator :

$$L_2 = \frac{h_2(8)}{h_2(8) + \cdots + h_5(8)} \tag{2.19}$$

The set of all individuals who are at risk at a given point in time is often referred to as the *risk set*. At month 8, the risk set consists of patient 2 through 5, inclusive.

We continue in this way for each successive event in order to construct each individual L_i. The general form is

$$L_i = \left[\frac{e^{\boldsymbol{\beta}\mathbf{x}_i}}{\sum_{j=1}^{N} y_{ij} e^{\boldsymbol{\beta}\mathbf{x}_j}} \right]^{\delta_i} \tag{2.20}$$

where $y_{ij} = 1$ if $t_j \geq t_i$ and $y_{ij} = 0$ if $t_j < t_i$ (the y's are just a convenient mechanism for excluding from the denominator those individuals who already experienced the event and are not part of the risk set). Moreover, the censored information are excluded because $\delta_i = 0$ for those cases. This expression is not valid for tied event times but it does allow for ties between event time and one or more censoring times.

A general expression for the partial likelihood for data with fixed covariates from a proportional hazards model is

$$PL = \prod_{i=1}^{N} \left[\frac{e^{\boldsymbol{\beta}\mathbf{x}_i}}{\sum_{j=1}^{N} y_{ij} e^{\boldsymbol{\beta}\mathbf{x}_j}} \right]^{\delta_i} \tag{2.21}$$

Once the partial likelihood is constructed, it can be maximized with respect to $\boldsymbol{\beta}$ just like an ordinary likelihood function. It is convenient to maximize the logarithm of the likelihood which is

$$\ln PL = \sum_{i=1}^{N} \delta_i \left[\boldsymbol{\beta}\mathbf{x}_i - \ln \left(\sum_{j=1}^{N} y_{ij} e^{\boldsymbol{\beta}\mathbf{x}_j} \right) \right] \tag{2.22}$$

Most partial likelihood programs use some version of the Newton-Raphson algorithm [Collett, 2003] to maximize this function with respect to $\boldsymbol{\beta}$.

The formula allowing to compute the standard error of the estimated parameter $\hat{\beta}$, are given in Appendix A of [Collett, 2003]. These standard errors can be used to obtain confidence intervals for β's. In particular, under the assumption that the estimated parameters $\hat{\beta}$'s follow a normal distribution, a $(100 - \alpha)\%$ confidence interval for a parameter β is the interval with limits $\hat{\beta} \pm z_{\alpha/2}\text{se}(\hat{\beta})$, where $z_{\alpha/2}$ is the upper $\alpha/2$-point of the standard normal distribution.

Normalization of the Loglikelihood We introduced a normalization for loglikelihood. Because the loglikelihood computed using (2.22) is directly proportional to the number of events, the normalized loglikelihood is simply the loglikelihood divided by the number of events in a dataset, given by

$$(normalized) \ln PL = \frac{1}{\sum_{i=1}^{N} \delta_i} \sum_{i=1}^{N} \delta_i \left[\boldsymbol{\beta}\mathbf{x}_i - \ln \left(\sum_{j=1}^{N} y_{ij} e^{\boldsymbol{\beta}\mathbf{x}_j} \right) \right] \tag{2.23}$$

Such a normalization is useful when we want to compare the loglikelihood of a model tested on different datasets. Indeed, these datasets may not contain the same number of events and the scales of the corresponding likelihoods may be very different. We will see in Section 5.1.4, the utility of this normalization.

2.4.2.2 Hypothesis Test

There exist three hypothesis tests in order to test the null hypothesis $H_0 : \beta = \beta^{(0)}$ where $\beta^{(0)}$ is the initial value for $\hat{\beta}$, the coefficients estimated by the Cox model. Only the Wald and the likelihood ratio tests will be described in this section[7].

- The Wald test is $(\hat{\beta} - \beta^{(0)})'\hat{\mathcal{I}}(\hat{\beta} - \beta^{(0)})$ where $\hat{\mathcal{I}} = \mathcal{I}(\hat{\beta})$ is the estimated information matrix[8] at the solution. For single variable, this reduces to the usual z-statistic $\hat{\beta}/\text{se}(\hat{\beta})$.
- The likelihood ratio test is $2\left(l(\hat{\beta}) - l(\beta^{(0)})\right)$ where l is the log partial likelihood at the initial and final estimates of $\hat{\beta}$.

The null hypothesis distribution of both the Wald and the likelihood ratio tests is a chi-square on p degrees of freedom where p is the number of coefficients. They are asymptotically equivalent but in finite samples they may differ. The likelihood ratio test is generally considered to be more reliable than the Wald test.

Such tests allow us to assess the likelihood that a coefficient or a set of coefficients in a Cox model are different from their initial values (typically 0).

We provide some additional topics about the semiparametric regression models in Appendix A. This concerns the treatment of tied data, the time-dependent covariates, the nonproportional hazards and the estimation of the survivor functions.

2.5 Testing for Differences in Survivor Functions

If a treatment has been applied to one group but not another, the obvious question to ask is “Did the treatment make a difference in the survival experience of the two groups ?”. Since the survivor function gives a complete accounting of the survival experience of each group, a natural approach to answering this question is to test the null hypothesis that the survivor functions are the same in the two groups : $S_1(t) = S_2(t)\ \forall t > 0$, where the subscripts distinguish the two groups.

There exist three alternative statistics for testing this null hypothesis : the logrank test (also known as the Mantel-Haenzel test), the Wilcoxon test and the hazard ratio.

[7] Details about the third hypothesis test, the score test, are given in [Therneau and Grambsch, 2000].

[8] The information matrix is the second derivative of the log partial likelihood with respect to β. Details are given in [Therneau and Grambsch, 2000].

2.5.1 Logrank Test

Suppose that there are r distinct event times, $t_1 < t_2 < \cdots < t_r$ across the two groups, and that at time t_j, d_{1j} individuals in group 1 and d_{2j} individuals in group 2 have an event occurrence, for $j = 1, 2, \ldots, r$. Suppose further that there are n_{1j} individuals at risk of event occurrence in the first group just before time t_j, and that there are n_{2j} at risk in the second group. Consequently, at time t_j, there are $d_j = d_{1j} + d_{2j}$ event occurrences in total out of $n_j = n_{1j} + n_{2j}$ individuals at risk. The situation is summarized in Table 2.4.

Group	Number of events at t_j	Number surviving beyond t_j	Number at risk just before t_j
1	d_{1j}	$n_{1j} - d_{1j}$	n_{1j}
2	d_{2j}	$n_{2j} - d_{2j}$	n_{2j}
Total	d_j	$n_j - d_j$	n_j

Table 2.4: Number of events at the jth event time in each of the two groups of individuals.

Each statistic can be written as a function of deviations of observed numbers of events from expected numbers. If the null hypothesis that survival is independent of group is true, we can therefore regard d_{1j}, the number of events at t_j in group 1, as the realization of a random variable D_{1j}, which can take any value in the range from 0 to $\min(d_j, n_{1j})$. In fact, D_{1j} has a distribution known as the *hypergeometric distribution* [Droesbeke, 1988], according to which the probability that D_{1j} in the first group takes the value d_{1j} is

$$\frac{\binom{d_j}{d_{1j}}\binom{n_j-d_j}{n_{1j}-d_{1j}}}{\binom{n_j}{n_{1j}}} \tag{2.24}$$

The mean of the hypergeometric random variable D_{1j} is given by

$$e_{1j} = \frac{n_{1j}d_j}{n_j} \tag{2.25}$$

so that e_{1j} is the expected number of individuals who have an event at time t_j in group 1.

For group 1, the logrank statistic can be written as

$$U_L = \sum_{j=1}^{r} (d_{1j} - e_{1j}) \tag{2.26}$$

Since the event times are independent of one another, the variance of (2.26) is simply the sum of the variances of the D_{1j}. D_{1j} having a hypergeometric distribution, the variance of D_{1j} is given by

$$\text{var}(D_{1j}) = \frac{n_{1j}(n_j - n1j)d_j(n_j - d_j)}{n_j^2(n_j - 1)} \tag{2.27}$$

so that the variance of U_L is

$$\text{var}(U_L) = \sum_{j=1}^{r} \text{var}(D_{1j}) = V_L \tag{2.28}$$

Furthermore, it can be shown that U_L has an approximate normal distribution when the number of event times is not too small [Droesbeke, 1988]. It then follows that $U_L/\sqrt{V_L}$ has a normal distribution with zero mean and unit variance. The square of a standard normal random variable has a chi-squared distribution of one degree of freedom (DF), denoted χ^2_1, and so we have that

$$\frac{U_L^2}{V_L} \sim \chi^2_1 \tag{2.29}$$

The p-value of the logrank test is calculated by using this chi-square statistic and a chi-square distribution with one DF.

2.5.2 Wilcoxon Test

The Wilcoxon statistic, given by

$$U_W = \sum_{j=1}^{r} n_j(d_{1j} - e_{1j}) \tag{2.30}$$

differs from the logrank test only by the presence of n_j, the total number at risk at each time point. Thus, it is a *weighted* sum of the deviations of observed numbers of events from expected numbers of events. As with the logrank statistic, the chi-square test is calculated by squaring the Wilcoxon statistic for either group and dividing by the estimated variance (see [Collett, 2003] for details).

Since the Wilcoxon test gives more weight to early times that to the late times (n_j always decreases), it is less sensitive than the logrank test to differences between groups that occur at later points in time. Although both statistics test the same null hypothesis, they differ in their sensitivity to various kinds of departures from that hypothesis. In particular, the logrank test is more powerful for detecting differences of the form

$$S_1(t) = [S_2(t)]^\gamma$$

where γ is some positive number other than 1. This equation defines a *proportional hazards model*, which is discussed in details in Section 2.4 (the logrank test is closely related to tests for differences between two groups that are performed within the framework of Cox's proportional hazards model). In contrast, the Wilcoxon test is more powerful than the logrank test in situations where event times have log-normal distributions with a common variance but with different means between the two groups. Neither test is particularly good when the survival curves cross [Allison, 1995].

The Wilcoxon and the logrank tests readily generalize to three or more groups, with the null hypothesis that all groups have the same survivor function. If the null hypothesis is true, all the test statistics have chi-square distributions with DF equal to the number of groups minus 1.

2.5.3 Hazard Ratio

The hazard ratio is a summary of the difference between two survival curves, representing the reduction in the risk of event between two different conditions. It is a form of relative risk. Proportional hazards regression model assumes that the relative risk of event between the two conditions is constant at each interval of time.

Let G be an indicator variable, which takes the value zero if an individual is on the first condition (e.g. low-risk group) and unity if an individual is on the second condition (e.g. high-risk group). If g_i is the value of G for the ith individual in the study, $i \in \{1, \ldots, N\}$, the hazard function for this individual can be written as

$$h_i(t) = \lambda_0(t) \exp(\beta g_i) \tag{2.31}$$

where $g_i = 1$ if the ith individual is on the second condition or zero otherwise. Because of the type of the indicator variable G, $\lambda_0(t)$ is the hazard function for an individual on the first condition. Moreover, the hazard function for any individual on the second condition is $\psi\lambda_0(t)$ (proportional hazards). ψ is the relative hazard or *hazard ratio* with $\psi = \exp(\beta)$

This is the proportional hazards model for the comparison of two groups. In this thesis, the indicator variable G is unity for the high-risk group and zero for the low-risk group. So the hazard ratio permits to assess if the risk of the high-risk group is higher than in the low-risk group.

Confidence Interval Once the parameter β is estimated, giving $\hat{\beta}$, the corresponding estimate of the hazard ratio is $\hat{\psi} = \exp(\hat{\beta})$, and the standard error of $\hat{\psi}$ can be obtained from the standard error of $\hat{\beta}$ (see Section 2.4.2.1). So, the standard error of $\hat{\psi}$ is given by

$$\mathrm{se}(\hat{\psi}) = \hat{\psi}\, \mathrm{se}(\hat{\beta}) \tag{2.32}$$

A $(100 - \alpha)\%$ confidence interval for the true hazard ratio ψ, can be obtained by exponentiating the confidence limit for β because the distribution of the logarithm of the estimated hazard ratio will be more closely approximated by a normal distribution than that of the hazard ratio itself [Collett, 2003].

2.6 Feature Selection

When a review of [Blum and Langley, 1997; Kohavi and John, 1997] on relevance including several papers on variable and feature selection was published, few studies used more than 40 features. The situation has changed considerably in the past few years and papers explore domains with hundreds to tens of thousands of variables or features. New techniques are proposed to address these challenging tasks involving many irrelevant and redundant variables and often comparably few training examples. Survival analysis of microarray data is such a new field with several thousands of genes for several hundreds of samples. Two characteristics of microarray data highlight the utility of feature selection :

- High feature/sample ratio : The microarray-based high-throughput technology generates a huge number of potential predictors (i.e. probes). On the other hand, the sample size of patients or cell lines is usually very small compared to the number of probes in the study (high feature/sample ratio). Modeling such high-dimensional data is complex. The problem becomes more difficult when the phenotypes such as time to death or time to cancer recurrence are subject to right-censoring. Additionally, microarray data often possess a great deal of noise.

  Due to the very high dimensional space of the predictors, the standard maximum Cox partial likelihood method cannot be applied directly to obtain the parameter estimates.

Moreover, from biological point of view, one should expect that only a small subset of the genes is relevant to predicting the phenotypes. Including all the genes in the predictive model increases its variance and is expected to lead to poor predictive performance.

- Highly correlated features : In microarray experiments, the expression levels of many probes may be highly correlated. Such a characteristic is explained by the co-regulation of many genes. Indeed, it has been assumed that similar patterns in gene expression profiles usually suggest relationships between the genes [Yu et al., 2003] or equivalently, the genes targeted by the same transcription factors tend to show similar expression patterns..

There are many potential benefits of variable and feature selection: facilitating data visualization and data understanding, reducing the measurement and storage requirements, reducing training and utilization times, defying the curse of dimensionality to improve prediction performance [Guyon and Elisseeff, 2003]. Some methods put more emphasis on one aspect or another, and this is another point of distinction between this special issue and previous work. Some papers focus mainly on constructing and selecting subsets of features that are useful to build a good predictor. This contrasts with the problem of finding or ranking all potentially relevant variables. Selecting the most relevant variables is usually suboptimal for building a predictor, particularly if the variables are redundant. Conversely, a subset of useful variables may exclude many redundant, but relevant, variables. For a discussion of relevance versus usefulness and definitions of the various notions of relevance, see the review articles [Blum and Langley, 1997; Kohavi and John, 1997].

Three aspects of feature selection will be tackled : filters that select variables by ranking them according to some statistic, subset selection methods including wrapper/embedded methods that assess subsets of variables according to their usefulness to a given predictor and feature construction which aim to increase the predictor performance by building more compact feature subsets.

2.6.1 Variable Ranking

Many variable selection algorithms include variable ranking as a principal or auxiliary selection mechanism because of its simplicity, scalability, and good empirical success. In many microarray studies [van't Veer et al., 2002; Chang et al., 2003; Jansen et al., 2005; Shipp et al., 2002], the gene ranking is a common method to select the most promising genes to build a classifier and/or to be the subject of further biological experiments. According to the design of the survival analysis, the ranking criterion is defined for individual variables, independently of the context of others.

Consider a set of N samples $(\mathbf{x}_i, y_i)$ $(i \in \{1, \ldots, N\})$ consisting of input values x_{ij} $(j \in \{1, \ldots, n\})$ and output value y_i . Variable ranking makes use of a scoring function $\mathcal{S}(j)$ associated to each input variable and computed from the values x_{ij} and y_i. By convention, we assume that a high score is indicative of a valuable variable and that we sort variables in decreasing order of $\mathcal{S}(j)$. To use variable ranking to build predictors, nested subsets incorporating progressively more and more variables of decreasing relevance are defined. A cross-validation procedure can be used to assess the optimal number of features [Vittinghof et al., 2005].

Following the classification of [Kohavi and John, 1997], variable ranking is a filter method : it is a preprocessing step, independent of the choice of the predictor. In practice, however, the scoring function is selected for its usefulness for a classifier that would use the most relevant input variables. So, the choice of the scoring function is not always independent on the choice of the classifier used in further analysis. Even if variable ranking is not optimal, it may be preferable to other variable subset selection methods because of its computational and statistical scalability. Computationally, it is efficient since it requires only the computation of n scores and the sorting of the scores. Statistically, it is robust against overfitting because it introduces bias but it may have considerably less variance [Hastie et al., 2001].

Three examples that outline the usefulness and the limitations of variable ranking techniques are given in [Guyon and Elisseeff, 2003].

2.6.2 Variable Subset Selection

Variable subset selection methods allow the selection of subsets of variables that together have good predictive power, as opposed to ranking variable methods that rank the variables according to their individual predictive power. We will focus on the wrappers which utilize the learning machine of interest as a black box to score subsets of variables according to their predictive power. Embedded methods which perform variable selection in the process of training and are usually specific to given learning machines, are considered as a promising approach for future works (see Section 6.1).

2.6.2.1 Wrappers and Embedded Methods

The *wrapper* methodology, recently popularized by [Kohavi and John, 1997], offers a simple and powerful way to address the problem of variable selection, regardless of the chosen learning machine. In fact, the learning machine is considered as a black box and the method is applicable to any learning algorithm, including off-the-shelf machine learning software packages. In its most general formulation, the wrapper methodology consists in using the prediction performance of a given learning machine to assess the relative usefulness of subsets of variables. In practice, one needs to define : (i) how to search the space of all possible variable subsets; (ii) how to assess the prediction performance of a learning machine to guide the search and halt it; and (iii) which learning machine to use. An exhaustive search can conceivably be performed, if the number of variables is not too large. But, the problem is known to be NP-hard [Amaldi and Kann, 1998] and the search becomes quickly computationally intractable. A wide range of search strategies can be used, including best-first, branch-and-bound, simulated annealing, genetic algorithms (see [Kohavi and John, 1997] for a review). Performance assessments are usually done using a validation set or by cross-validation.

Wrappers are often criticized because they seem to be a “brute force” method requiring massive amounts of computation, but it is not necessarily the case. Efficient search strategies may be devised. Using such strategies does not necessarily mean sacrificing prediction performance. In fact, it appears to be the inverse in some cases : greedy search strategies seem to be particularly computationally advantageous and robust against overfitting[9] [Guyon and Elisseeff, 2003]. Among such search strategies, we can mention two common methods : forward selection and backward elimination. In forward selection, variables are progressively incorporated into larger and larger subsets, whereas in backward elimination one starts with

[9] The name “greedy” come from the fact that one never revisits former decisions to include (or exclude) variables in light of new decisions.

the set of all variables and progressively eliminates the least promising ones. Both methods yield *nested subsets* of variables.

By using the learning machine as a black box, wrappers are remarkably universal and simple. But *embedded* methods that incorporate variable selection as part of the training process may be more efficient in several respects : they make better use of the available data by not needing to split the training data into a training and validation set; they reach a solution faster by avoiding retraining a predictor from scratch for every variable subset investigated. Recent articles highlight the promising results of embedded methods in the Cox regression as the LASSO procedure for Cox regression [Tibshirani, 1997] and penalized Cox regression [Gui and Li, 2004].

2.6.3 Feature Construction and Space Dimensionality Reduction

In some applications, reducing the dimensionality of the data by selecting a subset of the original variables may be advantageous for reasons including the expense of making, storing and processing measurements. If these considerations are not of concern, other means of space dimensionality reduction should also be considered.

The art of machine learning starts with the design of appropriate data representations. Better performance is often achieved using features derived from the original input. Building a feature representation is an opportunity to incorporate domain knowledge into the data and can be very application specific. Nonetheless, there are a number of generic feature construction methods, including: clustering; basic linear transforms of the input variables (e.g. PCA/SVD, LDA), etc (see [Dudoit et al., 2002] for a comparison of such methods).

Clustering has long been used for feature construction [Hartigan, 1975]. The idea is to replace a group of similar variables by a cluster centroid, which becomes a feature. The most popular algorithms include K-means and hierarchical clustering (see [Duda et al., 2001] for a review).

Clustering is usually associated with the idea of unsupervised learning (no use of any additional information such that demographic data) but it can be useful to introduce some supervision in the clustering procedure to obtain more discriminant features. This is the idea of the semi-supervised clustering [Bair and Tibshirani, 2004]. Firstly we rank the variables (see Section 2.6.1) using a supervised method (as Student t-test or univariate Cox model). Then we perform an unsupervised hierarchical clustering in order to cluster similar variables. Finally we use these clusters to construct new features.

2.6.3.1 Hierarchical Clustering

Hierarchical clustering is one of the most commmon clustering method [Hartigan, 1975]. [Eisen et al., 1998] introduced this method to analyze microarray data by organizing genes in a hierarchical tree structure (dendrogram), based on their degree of similarity. The basic idea is to assemble a set of items[10] into a binary tree, where items are joined by very short branches if they are very similar to each other, and by increasingly longer branches as their similarity decreases. A small example of such a tree is given in Figure 2.9.

Hierarchical clustering uses an agglomerative hierarchical processing consisting of repeated cycles where the two closest remaining items (those with the highest similarity) are joined by a node/branch of a tree, with the length of the branch set to the similarity between the joined

[10] An item refers to a variable or a set of variables.

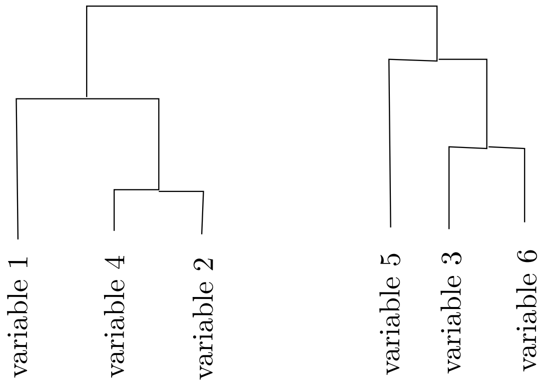


Figure 2.9: Example of tree computed by hierarchical clustering of six variables. Variable 4 and variable 2 are highly similar (joined by short branches). Idem for Variable 3 and variable 6 but with a smaller similarity between them, etc.

items. The two joined items are removed from list of items being processed and replaced by an item that represents the new branch. The similarities between this new item and all other remaining items are computed, and the process is repeated until only one item remains.

In order to apply this algorithm, we have to choose a metric of similarity and a way to compute the similarity between two items (called the linkage).

Metric of Similarity In this thesis, the uncentered Pearson correlation (sometimes called angular distance) is used as metric of similarity. The Pearson correlation (r) and the uncentered Pearson correlation (r_u) between two vectors $\mathbf{x}_{i_1}$ and $\mathbf{x}_{i_2}$, are given by (2.33) and (2.34) respectively

$$r(\mathbf{x}_{i_1}, \mathbf{x}_{i_2}) = \frac{\sum_{j=1}^{n} (x_{i_1j} - \bar{x}_{i_1})(x_{i_2j} - \bar{x}_{i_2})}{\sqrt{\sum_{j=1}^{n} (x_{i_1j} - \bar{x}_{i_1})^2 \sum_{j=1}^{n} (x_{i_2j} - \bar{x}_{i_2})^2}} \tag{2.33}$$

$$r_u(\mathbf{x}_{i_1}, \mathbf{x}_{i_2}) = \frac{\sum_{j=1}^{n} x_{i_1j} x_{i_2j}}{\sqrt{\sum_{j=1}^{n} x_{i_1j}^2 \sum_{j=1}^{n} x_{i_2j}^2}} \tag{2.34}$$

The Pearson correlation coefficient is always between -1 and 1, with 1 meaning that the two gene expression profiles are identical, 0 meaning they are completely uncorrelated, and -1 meaning they are perfect opposites. The correlation coefficient is invariant under linear transformation of the data [Droesbeke, 1988].

The uncentered version of the Pearson correlation coefficient differs in that there is no centering by subtraction of $\bar{x}_{i_1}$ and $\bar{x}_{i_2}$ from the expression measurements. So, two vectors differing only by an offset have $r = 1$ but $r_u \neq 1$. r_u is also called angular distance because it equals the cosine of the angle formed by the vectors $\mathbf{x}_{i_1}$ and $\mathbf{x}_{i_2}$.

Linkage There are a variety of ways to compute similarity between items that are sets of variables : centroid linkage, single linkage, complete linkage, and average linkage for instance (see [Duda et al., 2001]). In this thesis, the *complete linkage* is used to compute the similarity between two items c_1 and c_2 which is the maximum of all pairwise similarities between variables contained in c_1 and c_2. The Figure 2.10 gives a small example of the complete linkage.

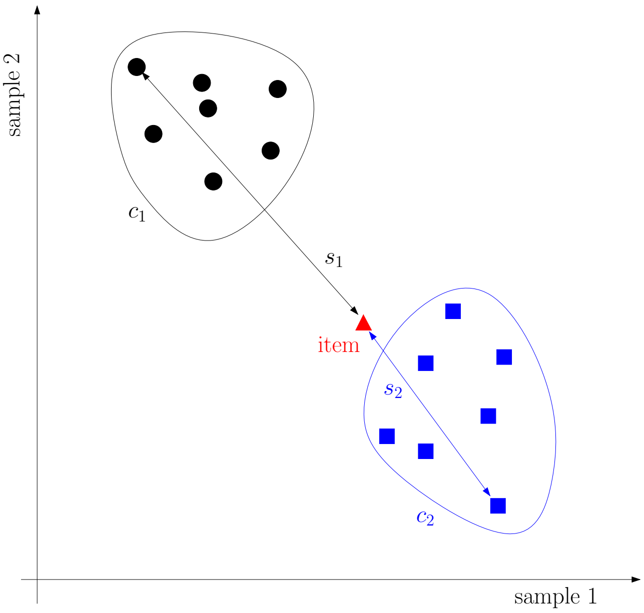


Figure 2.10: Example of complete linkage used in the hierarchical clustering with two clusters of variables c_1 and c_2 (circles and squares respectively) and only two samples. The new variable (triangle) will be assigned to cluster c_2 because $s_2 > s_1$ (a large similarity means a small distance between two probesets).

This method is computationally efficient. Indeed, a hierarchical clustering using the complete linkage computes only once the similarity matrix that contains all the pairwise similarities between probesets, and uses this information to construct the dendrogram.

Chapter 3

Materials

Contents

3.1 Populations

Three different populations of patients treated by TAMOXIFEN are studied in this thesis :

- OXFT : John Radcliffe Hospital (JRH), Oxford, UK (Dr Adrian Harris). 99 patients.
- KIT : Uppsala hospital (Karolinska) Sweden (Dr Jonas Bergh). Hybridized in Singapore (Lance Miller and Edison). 69 patients.
- GUYT : Guys Hospital, UK (Drs Paul Ellis and Cheryl Gillet). 87 patients.

All the patients had ER+ tumors, were under 88 years old and have been treated by TAMOXIFEN. Some patients have had to be discarded because of the insufficient follow-up and the lack of RNA material.

The Microarray Unit of Institut Jules Bordet (IJB) carried out the hybridizations for the OXFT and the GUYT populations. The hybridizations for the KIT population were performed by the laboratory of Lance Miller and Edison in Singapore. For the OXFT and KIT experiments, the chips HGU133A and HGU133B (see Section 3.2.1) have been hybridized. The new high density HGU133PLUS2 chip have been used for the GUYT population (see Section 3.2.1).

3.2 Microarray Platform

Microarray technology is a powerful tool for genetic research that uses nucleic acid hybridization techniques to evaluate the mRNA expression profile of thousands of genes within a single experiment. The Microarray Unit of the IJB uses the AFFYMETRIX© platform[1] which is

[1] `http://www.affymetrix.com`

a short oligonucleotide platform (see Appendix B for an overview of different microarray platforms). AFFYMETRIX© devices are shown in the Figure 3.1.

Figure 3.1: Affymetrix fluidic station and scanner.

3.2.1 Affymetrix© Technology

AFFYMETRIX© chips are short oligonucleotide (25 mers) arrays fabricated by direct synthesis of oligonucleotides on a silicon surface [Fdor et al., 1991]. Each chip contains up to 400,000 to 1M different probes (see Figure 3.2).

Since oligonucleotide probes are synthesized in known locations on the chip, the hybridization pattern and signal intensities can be interpreted in terms of gene identity and relative expression levels by a specific software[2]. Each gene is represented on the chip by a series of different pairs of oligonucleotide probes.

Each probe pair consists of a perfect match (called PM) and a mismatch (called MM) oligonucleotide (see Figure 3.3). The perfect match has a sequence exactly complementary to the particular region of gene and thus the probeset measures the expression of the gene. The mismatch probe differs from the perfect match probe by a single base substitution at the center base position, disturbing the bonding of the target gene transcript. This helps to determine the background and nonspecific hybridization (also called cross-hybridization) that contributes to the signal measured for the perfect match oligo [Lockhart et al., 1996]. Probes are selected on the basis of current information from GenBank and other nucleotide repositories. The sequences are believed to recognize unique regions of the 3' end of the gene.

The entire design of AFFYMETRIX© microarray experiments is depicted in Figure 3.4. Once the biological material under study is introduced in the chip, the hybridization process (see Figure 3.5) enables the assessment of the levels of expression of the genes characterized by the probes on the chips (see Figure 3.6).

Data Hierarchy As described previously, there are different levels of data in the AFFYMETRIX© technology :

1. The probes : the low-level measurements. The probes are constituted by two short oligonucleotides, the PM and the MM.

[2] We can mention the AFFYMETRIX© Microarray Suite Software or the Bioconductor [Gentleman et al., 2004] packages for R [R Development Core Team, 2005].

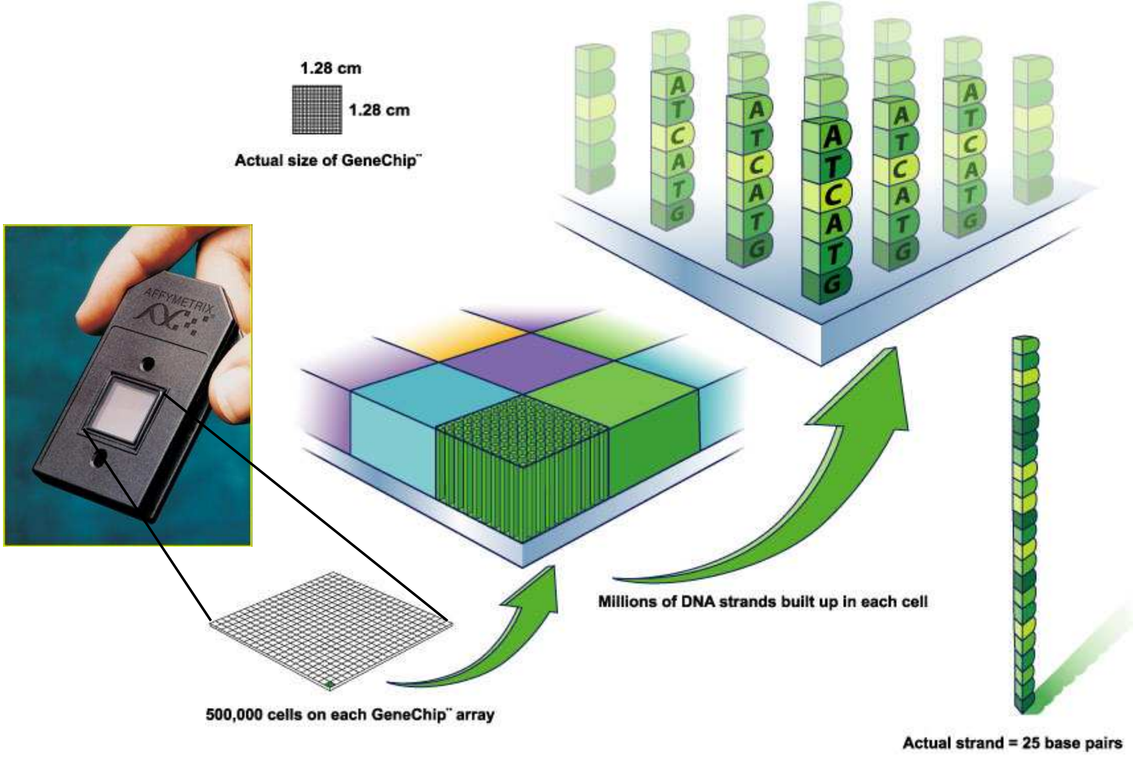

1.28 cm
1.28 cm
Actual size of GeneChip™
AFFYMETRIX
A
T
C
A
T
G
Millions of DNA strands built up in each cell
500,000 cells on each GeneChip™ array
Actual strand = 25 base pairs

Figure 3.2: Affymetrix© GeneChip probe array.

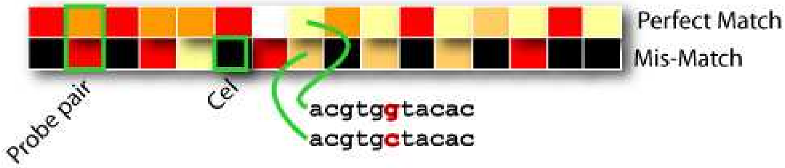


Figure 3.3: Oligonucleotide probe pair (Perfect Match and MisMatch).

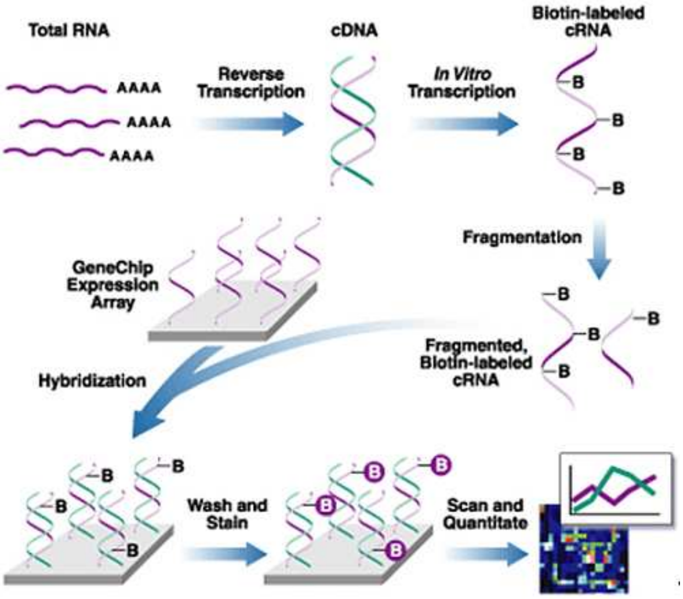


Figure 3.4: Design of AFFYMETRIX© microarray experiments.

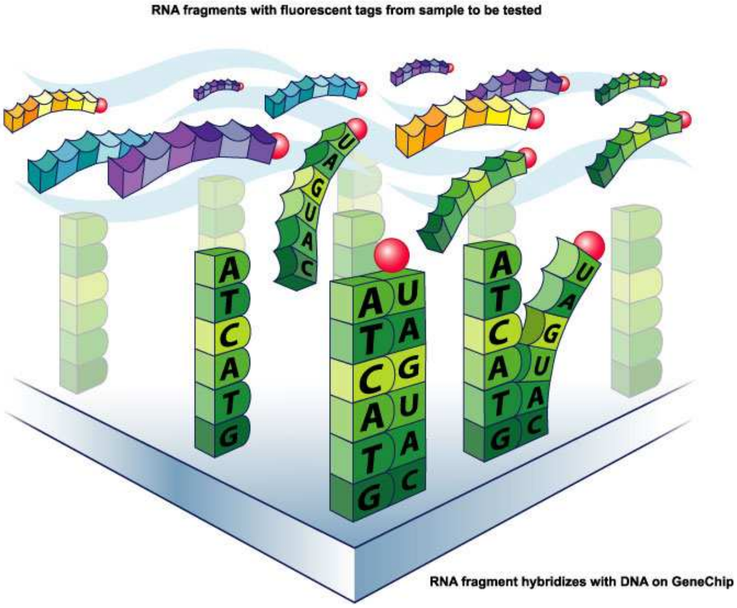

RNA fragments with fluorescent tags from sample to be tested
RNA fragment hybridizes with DNA on GeneChip
A
T
C
A
T
G
U
A
G
U
A
C

Figure 3.5: Hybridization process on the AFFYMETRIX© GeneChip array.

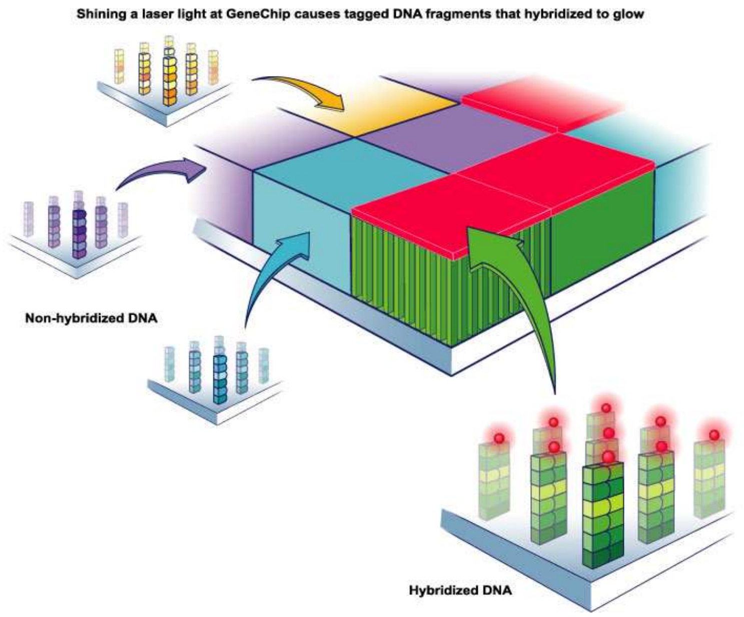

Shining a laser light at GeneChip causes tagged DNA fragments that hybridized to glow
Non-hybridized DNA
Hybridized DNA

Figure 3.6: Measurement of the level of gene expression after the hybridization process on AFFYMETRIX© GeneChip array.

2. The probesets : one probeset is a set of 11 to 20 probes.
3. The gene : one gene is represented by one or several probesets. The number of probesets depends on the sequence of the gene under study.

An example of such a hierarchy with a gene represented by two probesets, is depicted in Figure 3.7.

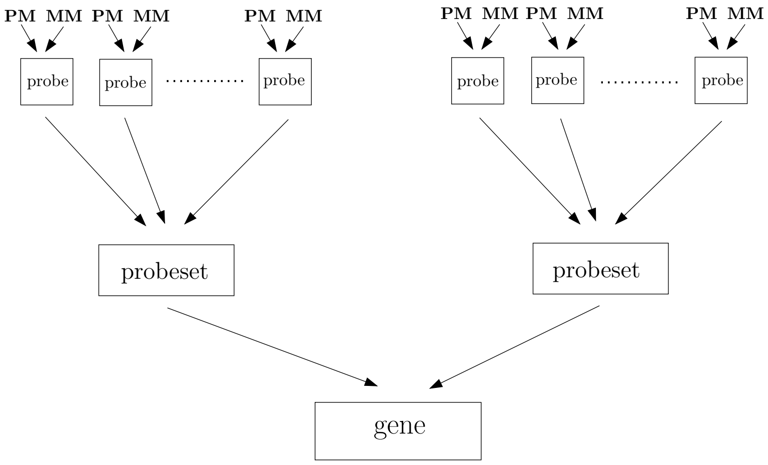


Figure 3.7: Data hierarchy on AFFYMETRIX© platform.

Affymetrix© Chips Several types of AFFYMETRIX© human chips for human samples are available : HGU95A, HGU95B, HGU133A, HGU133B, HGU133PLUS2, etc. For all the populations except the GUYT population, the samples were hybridized using the chips HGU133A

(22283 affy ids[3]) and HGU133B (22645 affy ids). The majority of known genes are on the chip HGU133A but the chip HGU133B is also used for the completeness (entire human genome). For the GUYT population, the new HGU133PLUS2 chip (54675 affy ids) is used. HGU133PLUS2 is the union of the chips HGU133A and HGU133B in a single high density chip. Its use permits to reduce the cost and the time of the experiments and the data can be compared with the data from the HGU133A and HGU133B chips.

[3] The majority of affy ids represent human genes but some are used for control or represent large region of transcribed DNA (EST).

Chapter 4

Methods

Contents

A flow-chart of our machine learning methodology is sketched in Figure 4.1. This methodology consists of the following steps :

1. Repartition of the dataset in training and test sets. Both sets are independent.

2. Training phase. This phase can be further decomposed in :

   Quality Assessment Quality assessment in order to discard microarray experiments that could have failed (see Section 4.1).

   Data Preprocessing The microarray data that have fulfilled the quality criteria, are preprocessed in order to obtain data comparable across samples (reading of the data and getting the expression measures) and to remove noninformative probesets (prefiltering) (see Section 4.2).

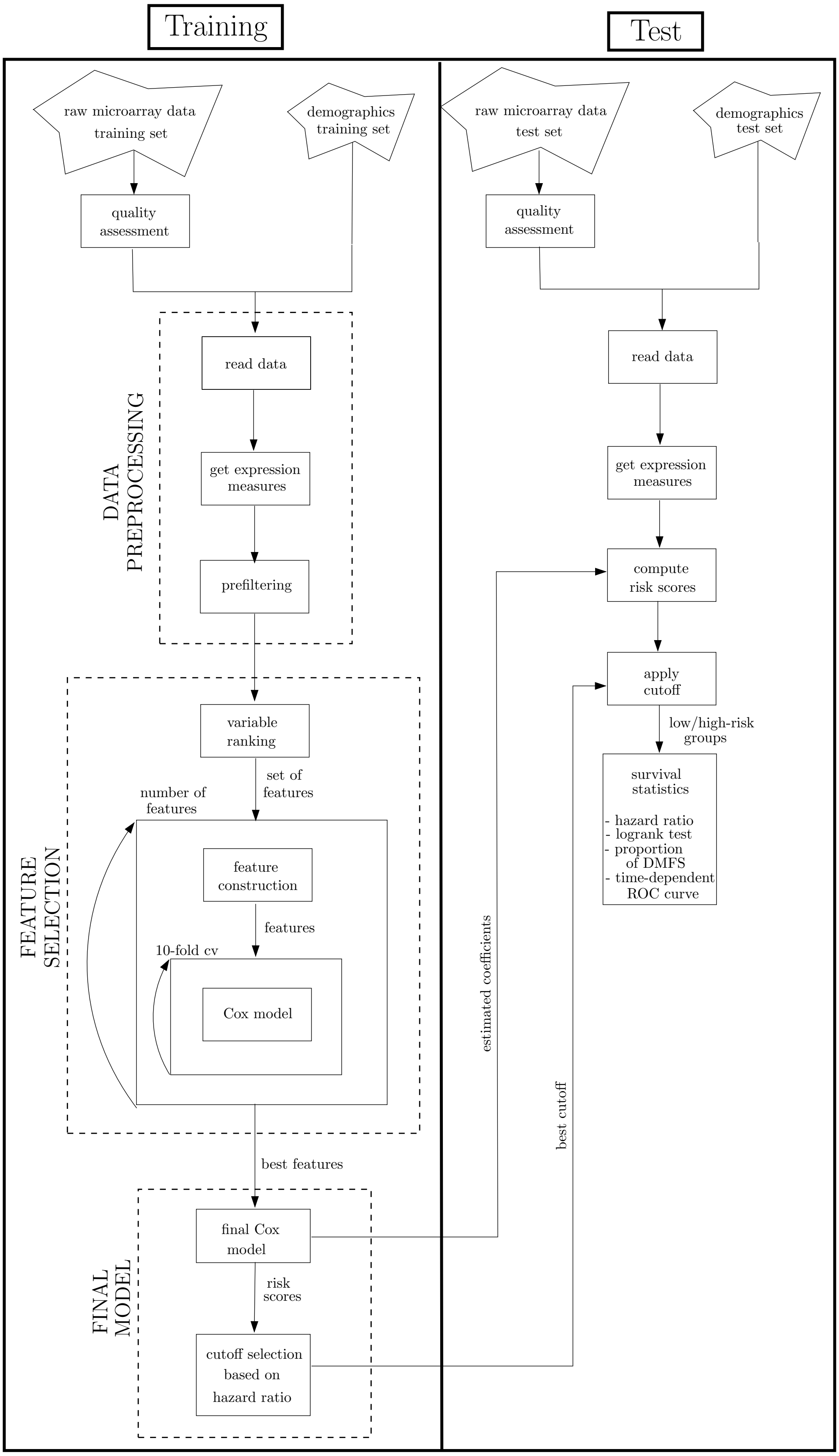


Figure 4.1: Design of the survival analysis.

Variable Ranking A ranking is performed on probesets in order to select the most promising ones for the final classifier (see Section 4.3.1).

Feature Construction From this subset of probesets, the features are constructed using a hierarchical clustering. In order to select the best number of clusters, a 10-fold cross-validation is performed to assess the performance of the classifier using such features (see Section 4.3.2).

Final Model Once the best set of features is constructed, a multivariate Cox model is fitted using all the training data to obtain the final model.

Risk Score Computation The final model is used to compute the risk score for each patient in training set.

Cutoff Selection In order to classify the patient in low and high-risk groups, a cutoff for the risk scores is selected based on survival statistics (see Section 4.4).

3. Test phase. This phase can be further decomposed in :

Quality Assessment Quality assessment in order to discard microarray experiments that could have failed (see Section 4.1).

Data Preprocessing The microarray data that have fulfilled the quality criteria, are preprocessed similarly to the training set.

Risk Score Computation The risk scores are computed for each patient in the test set using the final model fitted on the training set.

Apply Cutoff The same cutoff as selected in training part, is applied to classify the patients in low and high-risk groups.

Survival Statistics Assessment of the performance using the same survival statistics than those used in Section 4.4.

4.1 Quality Assessment

Depending on the microarray technologies, specific criteria have been used in literature to assess the quality of the microarray experiments. For the AFFYMETRIX© technology, two different sets of guidelines are commonly used in microarray studies : the AFFYMETRIX© guidelines [Affymetrix, 2002] and the Bioconductor[1] guidelines [Hartmann et al., 2003; Gautier et al., 2004]. A review of such guidelines is given in [Haibe-Kains, 2004].

4.2 Preprocessing Methods

We will describe the methods used in data preprocessing in this section. This procedure consists in reading the raw microarray data, in getting the expression measures and in performing a prefiltering of the probesets.

4.2.1 Read Data

The raw data are read using the functions of the *affy* package (see Bioconductor website[2] for the description of the *affy* package).

[1] See *affy* and *simpleaffy* packages.

[2] `http://www.bioconductor.org`

4.2.2 Get Expression Measures

The procedure used to get the expression measures of each probeset can be divided in four steps :

1. Background correction (B).
2. Normalization (N).
3. Summarization (S).
4. Population correction (P)

Let $\mathbf{x}$ be the raw intensities of a probeset coming from the CEL files of multiple microarray experiments (see Section 4.2.1). The expression measure s_c of this probeset (called *corrected signal*), is $s_c = P(S(N(B(\mathbf{x}))))$.

The *Robust Multi-array Average* procedure (RMA[3]) [Irizarry et al., 2003a] performs the first three steps. We introduce the population correction step described in Section 4.2.2.4.

4.2.2.1 Background Correction

Let us define the *background* as a measurement of signal intensity caused by auto-fluorescence of the microarray surface and cross-hybridization (see Section 3.2.1).

The background correction is a method which does some or all of the following :

- Corrects for background noise, biological sample preparation.
- Adjusts for cross-hybridization.
- Adjusts estimated expression values to fall on proper scale.

The RMA background correction is performed by estimating the unknown quantity S on the following model

$$O = S + \epsilon \tag{4.1}$$

where O is the observed PM intensity (see Section 3.2.1), S is the signal of interest and ϵ is a noise. S is assumed to have an exponential distribution with parameter α and ϵ is assumed to have a normal distribution with parameters μ (mean) and σ (standard deviation). To avoid any possibility of negative values, the normal is truncated at zero. Given we have o, the observed PM intensity, this then leads to an adjustment

$$E\left(s|O=o\right) = a + b\frac{\phi\left(\frac{a}{b}\right) - \phi\left(\frac{o-a}{b}\right)}{\Phi\left(\frac{a}{b}\right) - \Phi\left(\frac{o-a}{b}\right) - 1} \tag{4.2}$$

where $a = o - \mu - \sigma^2\alpha$ and $b = \sigma$. Note that ϕ and Φ are the normal PDF and the normal CDF respectively. α, μ and σ are then estimated [Irizarry et al., 2003b] what leads to the expected value of the signal, given the observed value of the intensity.

Note that the RMA procedure does not use the MM information (see Section 3.2.1) in order to correct signal for background and cross-hybridization. Indeed, exploratory analysis presented in [Naef et al., 2001; Irizarry et al., 2003b] suggests that the MM may be detecting signal as well as cross-hybridization and suggests to use only the PM information. So, only the PM intensities are corrected and used for further analysis.

[3] Currently, the RMA procedure is one of the most efficient as shown in [Bolstad et al., 2003].

4.2.2.2 Normalization

In microarray studies, biological sources of variation are referred to as *interesting variation*. However, many non-biological factors may contribute to the variability of data. This means that observed expression levels may also include variation introduced during the sample preparation, manufacture of the microarrays, and the processing of the microarrays (labeling, hybridization, and scanning). These are referred to as sources of *obscuring variation*. See [Hartemink et al., 2001; Irizarry et al., 2003b] for a more detailed discussion. The obscuring sources of variation can have many different effects on data. Unless microarrays are appropriately normalized, comparing data from different microarrays can lead to misleading results. Normalization is a process of reducing undesired variation across microarray experiments and may use information from multiple experiments.

Various methods have been proposed for normalizing AFFYMETRIX© GeneChip microarrays. [Bolstad et al., 2003] present a review of these methods and find *quantile* normalization to perform best. The aim of quantile normalization is to make the empirical distribution of probe intensities the same for all microarrays. So, the probe intensities distribution of sample i is identical to the probe intensities distribution of sample j with $i, j \in \{1, \ldots, N\}$.

Let $\mathbf{X}_{N \times P}$ be the matrix of the P probe intensities[4] for the N samples under study. The quantile normalization algorithm is given in algorithm 1.

Algorithm 1 Quantile normalization algorithm

```
Sort each line of $\mathbf{X}$ to give $\mathbf{X}_{sort}$ (save original positions).
Take the means across the columns of $\mathbf{X}_{sort}$ and assign this mean to each element in the
    column to give $\mathbf{X}^*_{sort}$.
Get $\mathbf{X}_{normalized}$ by rearranging each line of $\mathbf{X}^*_{sort}$ to have the same ordering as original $\mathbf{X}$
    (restore original positions).
```

The quantile normalization method is a specific case of the transformation $x_i^* = F^{-1}(G(x_i))$, where we estimate G by the empirical distribution of each microarray and F using the empirical distribution of the averaged sample quantiles.

4.2.2.3 Summarization

To obtain an expression measure of a probeset from intensities of the probes that belong to this probeset, we assume that for each probeset p, the background adjusted, normalized, and log transformed PM intensities, denoted with y, follow a linear additive model such that

$$y_{ipq} = \mu_{iq} + \alpha_{pq} + \epsilon_{ipq} \tag{4.3}$$

where $i \in \{1, \ldots, N\}$, $p \in \{1, \ldots, P\}$, $q \in \{1, \ldots, Q\}$, α_{jq} is the probe affinity effect of probeset q, μ_{iq} representing the log scale expression level of probeset q for array i, and ϵ_{ipq} representing an independent identically distributed error term with mean 0 of the probes intensities p belonging of the probeset q for the microarray i. For estimation of the parameters we assume that $\sum_{p=1}^{P} \alpha_{pq} = 0 \; \forall q \in Q$. This assumption is saying that AFFYMETRIX© has chosen probes with intensities that on average are representative of the associated genes

[4] Note that the background correction and the quantile normalization manage probe intensities. There are P probes in one microarray with $P \geq Q$ where Q is the number of probesets in one microarray (see Figure 3.7 for the data hierarchy).

expression. The estimate of μ_{iq} gives the expression measure for probeset q on microarray i. To protect against outlier probes a robust procedure is used, such as median polish [Holder et al., 2001; Tukey, 1977], to estimate model parameters.

4.2.2.4 Population Correction

In medical studies, it is common to study several distinct groups of patients, called populations. These populations come from different institutions[5] and their origin may be an important source of variability. We have observed that some probesets are systematically over-expressed or under-expressed according to the population of patients[6]. Unfortunately, the RMA procedure described above is not able to remove this source of variability without taking into account the origin of the samples (data not shown). Moreover, from a pragmatic point of view, it is easier to normalize each population separately and integrate the new samples when they are introduced in the analysis. In order to minimize the population effect, we introduce an additional transformation of the microarray data, called *population correction*.

Let $\mathbf{X}_{N \times Q}$ be the matrix of probeset expressions, after background correction, normalization and summarization for each population separately. Let *Pop* be the set of different populations, each population being a set of samples. The algorithm of the population correction method is given in algorithm 2.

Algorithm 2 Population correction algorithm

```
for each population $k \in Pop$ do
    for each probeset $q \in \{1, \dots, Q\}$ do
        $x_{median} \leftarrow$ median of probeset $q$ across the samples in $k$
        for each sample $i \in k$ do
            $\mathbf{X}^*[i, q] \leftarrow \mathbf{X}[i, q] - x_{median}$          ▷ median centering
        end for
    end for
end for
$\mathbf{X}^*[i, q]$ contains the corrected probeset intensities.
```

In other words, each probeset is centered by its median in taking into account the origin of the samples (population). After the population correction, we have no more observed[7] the over/under-expression of the probesets of interest and the patients of different populations can be compared for further analysis. This correction has already been applied with success in [Sotiriou et al., 2005].

4.2.3 Prefiltering

The prefiltering consists in removing some noninformative probesets without using demographic information. The prefiltering is made of two steps :

- Remove the AFFYMETRIX© control probesets (see [Affymetrix, 2002]).

[5] In the TAMOXIFEN resistance project, we have three different populations of patients (see Section 3.1).

[6] A hierarchical clustering (see Section 2.6.3.1) can highlight the population effect. Indeed if you cluster all your experiments and you observe that your experiments from the same population are cluster together, that means that the population effect is stronger that biological information.

[7] A hierarchical clustering (see Section 2.6.3.1) can no more highlight the population effect.

- Remove the probesets that have at least 95% of *Absent calls* among all the samples in the training set. The default parameters are used in the detection calls method.

Detection Calls A detection algorithm [Affymetrix, 2002] uses probe pair (PM/MM, see Section 3.2.1) intensities to generate detection p-value and to assign a *Present*, *Marginal*, or *Absent* call to each probeset. Each probe pair in a probeset is considered as having a potential vote in determining whether the measured transcript is detected (Present) or not detected (Absent). The vote is described by a value called the *discrimination score* (R). The score is calculated for each probe pair and is compared to a predefined threshold τ. Probe pairs with scores higher than τ vote for the presence of the transcript and inversely. The voting result is summarized as a p-value associated with the test of the difference between score and τ.

The discrimination score is a basic property of a probe pair that describes its ability to detect its attended target (see Figure 4.2)

$$R_{ip} = \frac{(\mathbf{X}_{pm}[i,p] - \mathbf{X}_{mm}[i,p])}{(\mathbf{X}_{pm}[i,p] + \mathbf{X}_{mm}[i,p])} \tag{4.4}$$

where $i \in \{1,\ldots,N\}$, $p \in \{1,\ldots,P\}$, *pm* and *mm* indexes indicating the PM and the MM intensity of the probe respectively.

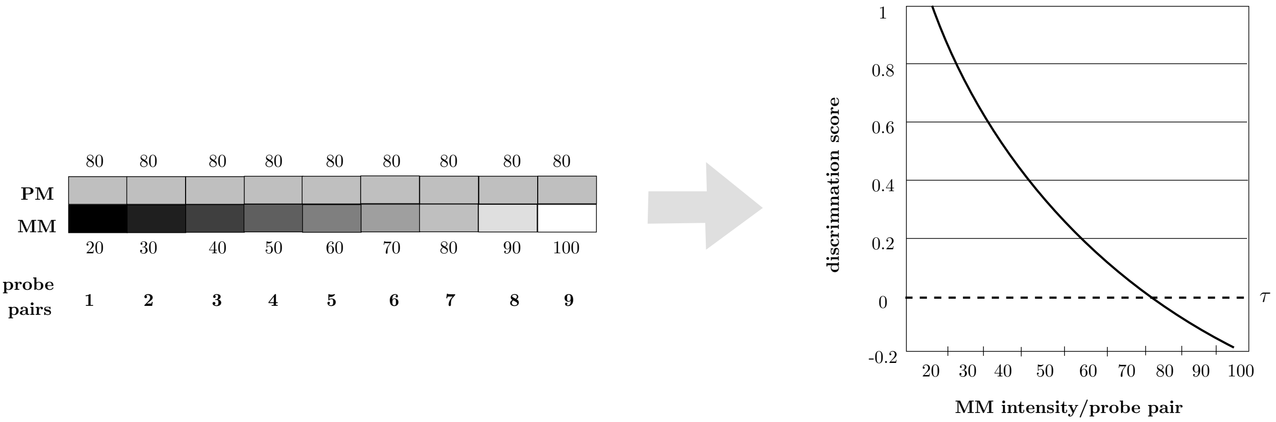


Figure 4.2: Example of discrimination score [Affymetrix, 2002]. The PM intensity is fixed to 80 and the MM intensity varies from 10 to 100. The y-axis represents the discriminant score and the x-axis represents the MM intensity.

Each discrimination score is compared to the threshold τ, which is a small positive number[8] that can be adjusted to increase or decrease the sensitivity and/or specificity of the analysis. Detection p-value is calculated by the One-Sided Wilcoxon's Signed Rank test [Droesbeke, 1988]. Finally, a detection call *Present/Marginal/Absent* is assigned to each probeset according to its detection p-value and two arbitrary thresholds α_1 and α_2 (see Figure 4.3).

4.3 Feature Selection

The feature selection step aims to identify a set of features giving good performance in generalization when used in the classifier. These methods includes the variable ranking, the feature

[8] Default value equals to 0.015.

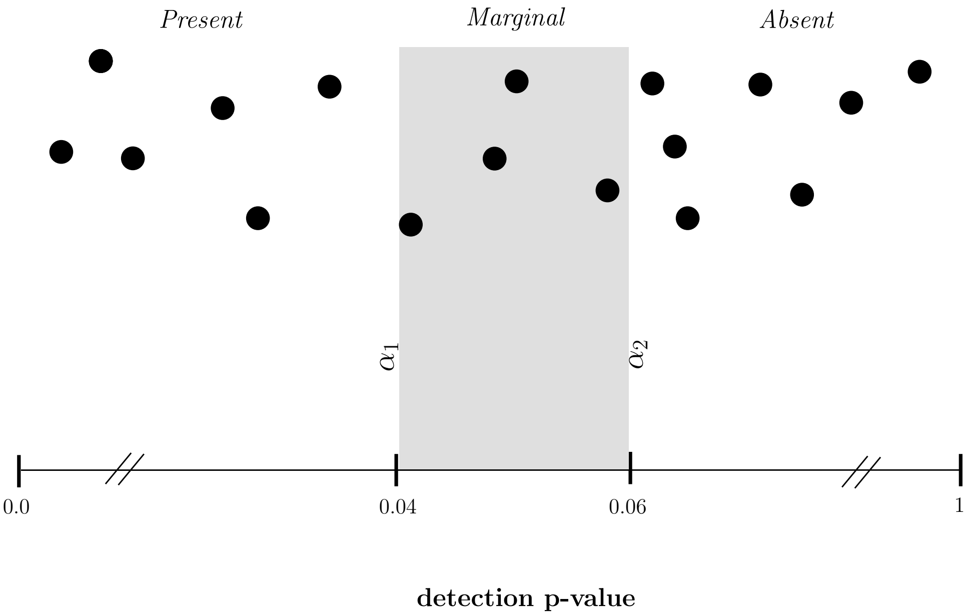


Figure 4.3: Detection p-value [Affymetrix, 2002].

construction and the Cox model as classifier.

In the following sections, by “variables” we refer to the “raw” input variables (probeset expression obtained after the preprocessing step, see Section 4.2) and by “features” we refer to the variables constructed from input variables.

4.3.1 Variable Ranking

For an efficient variable ranking (see Section 2.6), we have to choose a scoring function useful for the fitting of the classifier. Because the model used for classification of patients is a multivariate Cox model, we have selected a scoring function based on a univariate Cox model. Many other alternatives can be considered like a variable ranking based on Student t-test or Pearson correlation.

4.3.1.1 Scoring Function Based on Univariate Cox Model

The scoring function $\mathcal{S}(j)$ is one minus the p-value computed by a likelihood ratio test (see Section 2.4.2.2) of a univariate Cox model. The p-value for the variable j is computed from the χ^2 distribution using the following statistic :

$$\begin{aligned} \chi^2 \; statistic(j) &= 2\left(l(\hat{\beta}^{(j)}) - l(\beta^{(0)})\right) \\ &= 2\left(\sum_{i=1}^{N} \delta_i \left[\hat{\boldsymbol{\beta}}^{(i)}\mathbf{x}_{ij} - \ln\left(\sum_{k=1}^{N} y_{ik} e^{\hat{\boldsymbol{\beta}}^{(j)}\mathbf{x}_{kj}}\right) + \ln\left(\sum_{k=1}^{N} y_{ik}\right)\right]\right) \end{aligned}$$

where $y_{ik} = 1$ if $t_k \geq t_i$ and $y_{ik} = 0$ if $t_k < t_i$, $\hat{\beta}^{(j)}$ is the vector of the estimated coefficient of the variable j and $\beta^{(0)}$ represents the null coefficient.

The p-value of the likelihood ratio test represents the significance of the difference between the partial loglikelihoods of the models with and without the considered variable. In other words, how much the variable is valuable for the model.

4.3.2 Feature Construction

The feature construction step aims at constructing features derived from the original input (see Section 2.6.3). In the survival analysis design given in Figure 4.1, we use such a method to construct the features :

- We perform a variable ranking using the scoring function described in the previous section and we choose a threshold to select only the informative probesets (e.g. only the probesets that have a score > 0.9999).
- A hierarchical clustering (see Section 2.6.3.1) is performed in order to select clusters of highly correlated probesets.
- For each cluster, new features are constructed in computing the cluster centroid, i.e. average of the intensities of all the probesets in a cluster.

This semi-supervised method (see Section 2.6.3) selects the probesets using demographic data (supervised) and a hierarchical clustering (unsupervised) is used to cluster highly correlated probesets to obtain cluster centroids.

Such a method has several advantages :

1. Variable ranking is less prone to overfitting (see Section 2.6.1) and is computationally efficient. Indeed we have to deal with more than 30,000 probesets.
2. Clustering of highly correlated probesets[9] allows us to identify interesting groups of co-regulated genes which will be the object of further biological experiments.
3. The computation of the cluster centroids permits (i) to reduce the variance of the features (ii) to facilitate the validation of the classifier on another microarray platform (see Section 4.3.2.1).

We use a hierarchical clustering with uncentered Pearson correlation as similarity metric and complete linkage (see Section 2.6.3.1).

4.3.2.1 Classifier Validation on Different Microarray Platforms

In this section, we propose a method to facilitate the validation of a classifier developed on a specific microarray platform (e.g. AFFYMETRIX©) to a different one (e.g. AGILENT©). Because of the heterogeneity of the existing microarray platforms, it is very hard to compare/validate new results between different microarray studies. However, given the cost and the scarcity of microarray experiments, it would be very interesting to be able to test the final model of classification on other microarray data.

There are two main difficulties with such comparisons/validations :

[9] A cluster of probesets can be reduced to a cluster of genes in consulting their biological annotations (see Appendix D).

1. We have to find similar probes on the two microarray platforms under study (e.g. a probe representing the same gene in AFFYMETRIX© and AGILENT© platforms).

2. We have to normalize the datasets coming from the different microarray platforms in order to analyze comparable data.

The first problem is partly solved by the design of the feature selection (see Section 4.3). Indeed, the final classification model is based on a multivariate Cox regression fitted with the constructed features. The set of features is composed by the cluster centroids constructed during the feature construction step. Because each feature is an average of several probesets in a cluster, the classifier is less sensitive to the absence of one or more probes as it may happen when you analyze data coming from different microarray platforms. The robustness of the classifier according to the loss of one or more probes will be analyzed in future works (see Section 6.1).

4.3.3 Cox Model

Once the features are constructed, the normalized loglikelihood (see equation (2.23)) of a multivariate Cox model fitted using these features is estimated by a 10-fold cross-validation procedure. This procedure partitions the training data in ten couples $\{training\ subset, test\ subset\}$ where the training subset contains 90% of the training set and the test subset contains the remaining 10%. For each couple, a multivariate Cox model is fitted using the training subset and its loglikelihood is computed on the corresponding test subset and normalized.

At the end, the ten normalized loglikelihoods are averaged to obtain an estimation of the classifier performance on independent data. Such a procedure is represented in Figure 4.4.

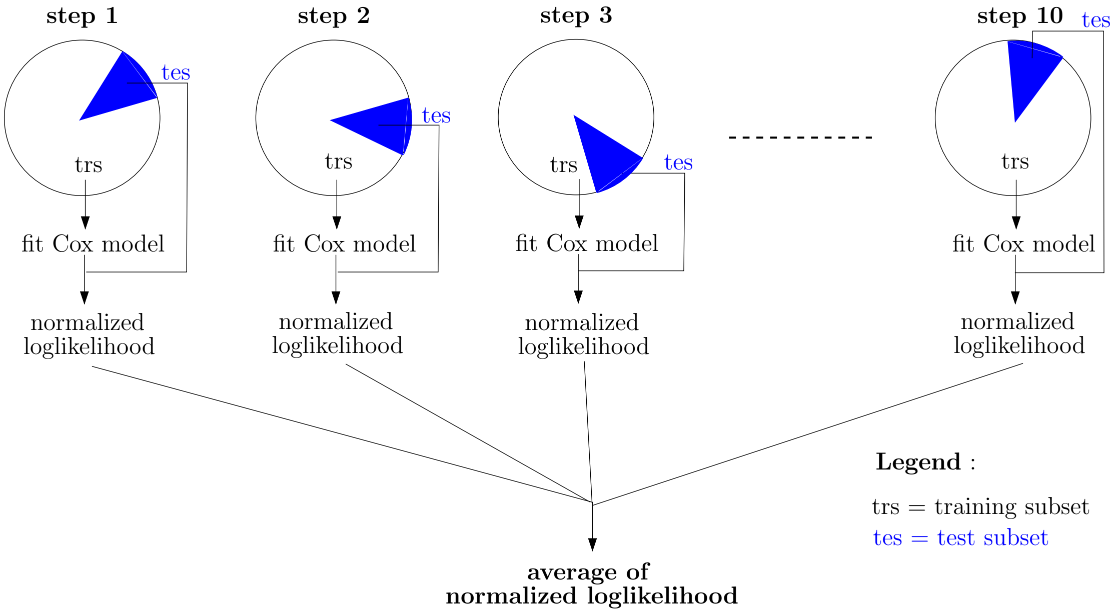


Figure 4.4: 10-fold cross-validation procedure to estimate the loglikelihood of the Cox model on independent data.

4.4 Final Model

The last part of the analysis concerns the fitting of the final Cox model and its use to classify the patients in low and high-risk groups. In order to assess the difference in survival between the two groups of patients, we compute several survival statistics.

4.4.1 Final Cox Model

Once the best set of constructed features is selected, a Cox model is fitted using all the training set. This model is used to compute the risk scores rs_i such that

$$rs_i = \sum_{j=1}^{F} \hat{\beta}_j f_{ij} \tag{4.5}$$

where F is the number of features, $i \in \{1, \dots, N\}$, $j \in \{1, \dots, F\}$, $\hat{\beta}$'s are the estimated coefficents of the final Cox model and f_{ij} is the feature j of the sample i. The construction of such features is described in Section 4.3.2.

4.4.2 Cutoff Selection

The risk score rs is a continuous variable representing the risk for each patient to die. In order to classify the patient in two groups, a cutoff has to be selected. The aim is to have two groups of patients with high difference in survival. Such a difference can be assessed with different kind of survival statistics (see Section 4.5). The cutoff for the risk scores is selected on the basis of the hazard ratio (HR, see Section 2.5.3).

The algorithm of cutoff selection based on hazard ratio is given in algorithm 3.

Algorithm 3 Algorithm of the cutoff selection based on hazard ratio

```
Consider only the patients on the training set.
Keep only cutoffs which leave at least 25% of patients in the high-risk group.
Keep only cutoffs which have not the unity in their 95% confidence interval (see Section
  4.5.1).          ▷ $HR = 1$ means no difference in survival between low and high-risk groups
Select the cutoff which has the lowest proportion of DMFS at 3 years (see Section 4.5.3)
  and the highest HR.
```

4.5 Survival Statistics

Several ways to assess the difference in survival between low and high-risk groups are described in this section.

4.5.1 Hazard Ratio

The hazard ratio was introduced in Section 2.5.3. This statistic permits to assess the reduction in the risk of event between two different groups.

4.5.2 Logrank Test

The logrank test was introduced in Section 2.5. This statistical test permits to test the difference between two survivor functions, S_1 and S_2. They are estimated using the KM estimator (see Section 2.3.1) from the patients belonging to low and high-risk groups.

4.5.3 Proportion of DMFS

The survival at three and five years are common criteria in clinical practice. In this thesis, this criterion is called the *distant metastasis free survival* (DMFS) because we study the appearance of distant metastases for the patients under a specific treatment. It would be interesting to have a classifier specially efficient to discriminate patients who would not die in the first three of five years (low-risk group) even if the high-risk is less well discriminated (trade-off between sensitivity and specificity). The same reasoning can be made for the inverse.

The proportion of events (distant metastases) during the first three or five years, is computed for low and high-risk groups in order to assess such a performance criterion for the classifier.

4.5.4 Time-Dependent ROC Curve

ROC curves are a popular method for displaying sensitivity and specificity of a continuous diagnostic marker R, for a binary disease variable D. However, many disease outcomes are time-dependent $D(t)$, and ROC curves that vary as a function of time may be more appropriate. A common example of a time-dependent variable is vital status, where $D(t) = 1$ if a patient has died prior to time t and zero otherwise. In [Heagerty et al., 2000], the authors propose summarizing the discrimination potential of a marker R, measured at baseline $t = 0$, by calculating ROC curves for cumulative disease or death incidence by time t, which is denoted as $ROC(t)$.

A typical complexity with survival data is that observations may be censored. Two ROC curve estimators are proposed that can accommodate censored data. A simple estimator is based on using the Kaplan-Meier estimator for each possible subset $R > c$. However, this estimator does not guarantee the necessary condition that sensitivity and specificity are monotone in R. An alternative estimator that does guarantee monotonicity is based on a nearest neighbor estimator for the bivariate distribution function of (R, T), where T represents survival time [Akritas, 1994]. Two interesting examples are given in [Heagerty et al., 2000][10].

4.5.4.1 Sensitivity and Specificity

The sensitivity and the specificity of a binary classification test or algorithm are parameters that express something about the test's performance. The sensitivity of such a test is the proportion of those cases having a positive test result of all positive cases tested ($\frac{TP}{TP+FN}$[11]). The specificity of such a test is the proportion of true negatives of all the negative samples

[10] In [Heagerty et al., 2000], the authors present an example where $ROC(t)$ is used to compare a standard and a modified flow cytometry measurement for predicting survival after detection of breast cancer and an example where the $ROC(t)$ curve displays the impact of modifying eligibility criteria for sample size and power in HIV prevention trials.

[11] TP, FN, TN, FP represent the rate of true positives, false negatives, true negatives and false positives respectively.

tested ($\frac{TN}{TN+FP}$). Sensitivity and specificity are well established for simple binary variables with either discrete or continuous marker measurements. In [Heagerty et al., 2000], this concepts of sensitivity and specificity are extended to time-dependent binary variables such as vital status, allowing characterization of diagnostic accuracy for censored data.

For test results defined on continuous scales, ROC curves are standard summaries of accuracy. If R denotes the diagnostic test or marker, with higher values more indicative of disease, and D is a binary indicator of disease status, then the ROC curve for R is a plot of the sensitivity associated with the dichotomized test $R > c$ versus $(1 - specificity)$ for all possible threshold values c, i.e. the ROC curve is the monotone increasing function in $[0, 1]$ with coordinates $(\Pr\{R > c \mid D = 0\}, \Pr\{R > c \mid D = 1\})$ where $c \in \{-\infty, \infty\}$. This function characterizes the diagnostic potential of a continuous test by summarizing all of the possible trade-offs between sensitivity and specificity. The higher the ROC curve is in the quadrant $[0, 1] \times [0, 1]$, the better is its capacity for discriminating diseased from nondiseased subjects.

Definitions Let T_i denote failure time and R_i the diagnostic marker for subject i. Let C_i denote the censoring time, $Z_i = min(T_i, C_i)$ the follow-up time, and δ_i a censoring indicator with $\delta_i = 1$ if $T_i \leq C_i$ and $\delta_i = 0$ if $T_i > C_i$. We use the counting process $D_i(t) = 1$ if $T_i \leq t$ and $D_i(t) = 0$ if $T_i > t$ to denote event (disease) status at any time t with $D_i(t) = 1$ indicating that subject i has had an event prior to time t.

Recall that ROC curves display the relationship between a diagnostic marker R, and a binary disease variable D_i by plotting estimates of the sensitivity $\Pr\{R > c \mid D = 1\}$, and one minus the specificity $1 - \Pr\{R \leq c \mid D = 0\}$ for all possible values c. When disease status is time dependent, consider sensitivity and specificity as time-dependent functions and define them as

$$\begin{aligned} sensitivity(c, t) &= \Pr\{R > c \mid D(t) = 1\} \\ specificity(c, t) &= \Pr\{R \leq c \mid D(t) = 0\} \end{aligned}$$

Using these definitions, we can define the corresponding ROC curve for any time t, $ROC(t)$.

Kaplan-Meier Estimator We can use Bayes' theorem to rewrite the sensitivity and the specificity as

$$\begin{aligned} \Pr\{R > c \mid D(t) = 1\} &= \frac{\{1 - S(t \mid R > c)\} \Pr\{R > c\}}{1 - S(t)} \\ \Pr\{R \leq c \mid D(t) = 0\} &= \frac{S(t \mid R \leq c) \Pr\{R \leq c\}}{S(t)} \end{aligned}$$

where $S(t)$ is the survival function $S(t) = \Pr\{T > t\}$ and $S(t \mid R > c)$ is the conditional survival function for the subset defined by $R > c$.

A widely used nonparametric estimate of $S(t)$ is given by the KM estimator [Kaplan and Meier, 1958] (see Section 2.3.1). The KM estimator uses all the information in the data, including censored observations, to estimate the survival function.

A simple estimator for sensitivity and specificity at time t is then given by combining the

KM estimator $\widehat{S}_{KM}(t)$ and the empirical distribution function of the marker covariate R, as

$$\begin{aligned}\widehat{\Pr}_{KM}\{R > c \,|\, D(t) = 1\} &= \frac{\{1 - \widehat{S}_{KM}(t \,|\, R > c)\}\{1 - \widehat{F}_R(c)\}}{\{1 - \widehat{S}_{KM}(t)\}} \\ \widehat{\Pr}_{KM}\{R \leq c \,|\, D(t) = 0\} &= \frac{\widehat{S}_{KM}(t \,|\, R \leq c)\widehat{F}_R(c)}{\widehat{S}_{KM}(t)}\end{aligned}$$

where $\widehat{F}_R(c) = \dfrac{\sum \mathbf{1}(R_i \leq c)}{n}$.

Now, we can estimate sensitivity and specificity for a linear predictor with censored data using the KM estimator. However this estimator has two problems :

1. This simple estimator does not guarantee that sensitivity or specificity is monotone. By definition, we gave $\Pr\{R > c \,|\, D(t) = 1\} \geq \Pr\{R > c' \,|\, D(t) = 1\}$ for $c' > c$. See a violation example in [Heagerty et al., 2000].

2. A potential problem with the KM-based ROC estimator is that the conditional KM estimator $\widehat{S}_{KM}(t \,|\, R > c)$ assumes that the censoring process does not depend on R. This assumption may be violated in practice when the intensity of follow-up efforts are influenced by the baseline diagnostic marker measurements.

For the moment, we have implemented only the KM-based ROC estimator.

4.5.4.2 Area Under the ROC Curve

The *area under the ROC curve* (AUC) can be interpreted as the probability that the test result from a randomly chosen diseased individual exceeds that for a randomly chosen nondiseased individual and is often used to summarize the ROC curve.

Chapter 5

Results

Contents

In order to study the effectiveness of the methods described in Chapter 4 on real data, we apply the design of the survival analysis given in Figure 4.1 to the data coming from the Tamoxifen resistance project (see Sections 1.1.1.1).

We use the OXFT population (99 patients) as training set and the KIT and GUYT populations as test set (156 patients). The Table 5.1 gives a summary of the repartition of the data and their main characteristics.

Tamoxifen resistance project	Training set	Test set
Populations	OXFT	KIT and GUYT
Number of patients	99	156
Number of probesets	44928	44928

Table 5.1: Repartition of the data between training and test sets for the Tamoxifen resistance project.

5.1 Tamoxifen Resistance Project

5.1.1 Quality Assessment

The 255 microarrays coming from the OXFT, KIT and GUYT populations have passed the quality tests described in Section 4.1 (data not shown due to confidentiality).

5.1.2 Preprocessing Methods

We apply the preprocessing methods described in Section 4.2 to the data, resulting in 32,139 probesets (we discard 12,789 probesets at the prefiltering step, see Section 4.2.3).

5.1.3 Variable Ranking

The scoring function (see Section 4.3.1.1) is applied to all prefiltered probesets. The histogram in Figure 5.1 gives an approximate of the score distribution. We can see that there is a large number of very high scores (close to 1) in comparison to smaller scores (0 to 0.9). In order to keep only a small subset of promising probesets, we have to choose a large threshold. This results in 213 probesets with a score > 0.9999.

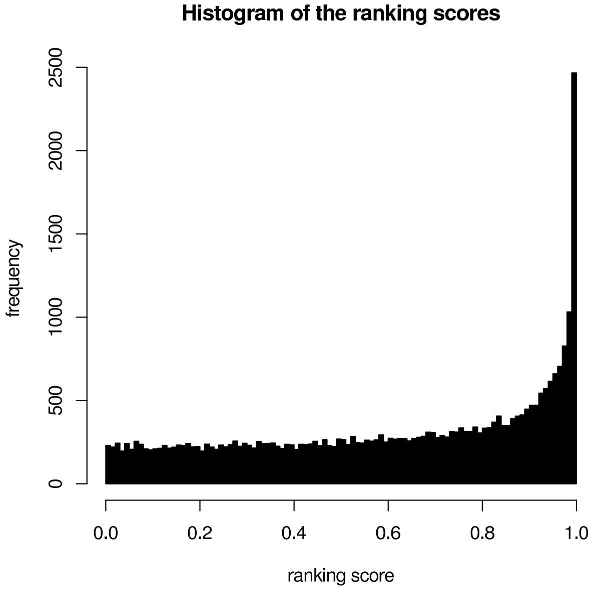


Figure 5.1: Histogram of the score computed by the scoring function for all the probesets remaining after the prefiltering.

The annotations of the remaining probesets are given in Appendix C. These annotations are available using the *annotation* package of Bioconductora and the information publicly available on the AFFYMETRIX© website[1].

5.1.4 Feature Construction

In this step, we carry out hierarchical clustering (see Section 2.6.3.1) on the probesets selected after variable ranking in order to cluster probesets according to a correlation metric. The resulting clustering is given in Figure 5.2. This figure includes the dendrogram (tree at the top) and the *heatmap* (below the dendrogram). The heatmap is a graphical representation of the gene expressions with down-regulation in green (negative gene expression), up-regulation in red (positive gene expression) and absence of expression in black (gene expression close to zero).

[1] `http://www.affymetrix.com/analysis/index.affx`

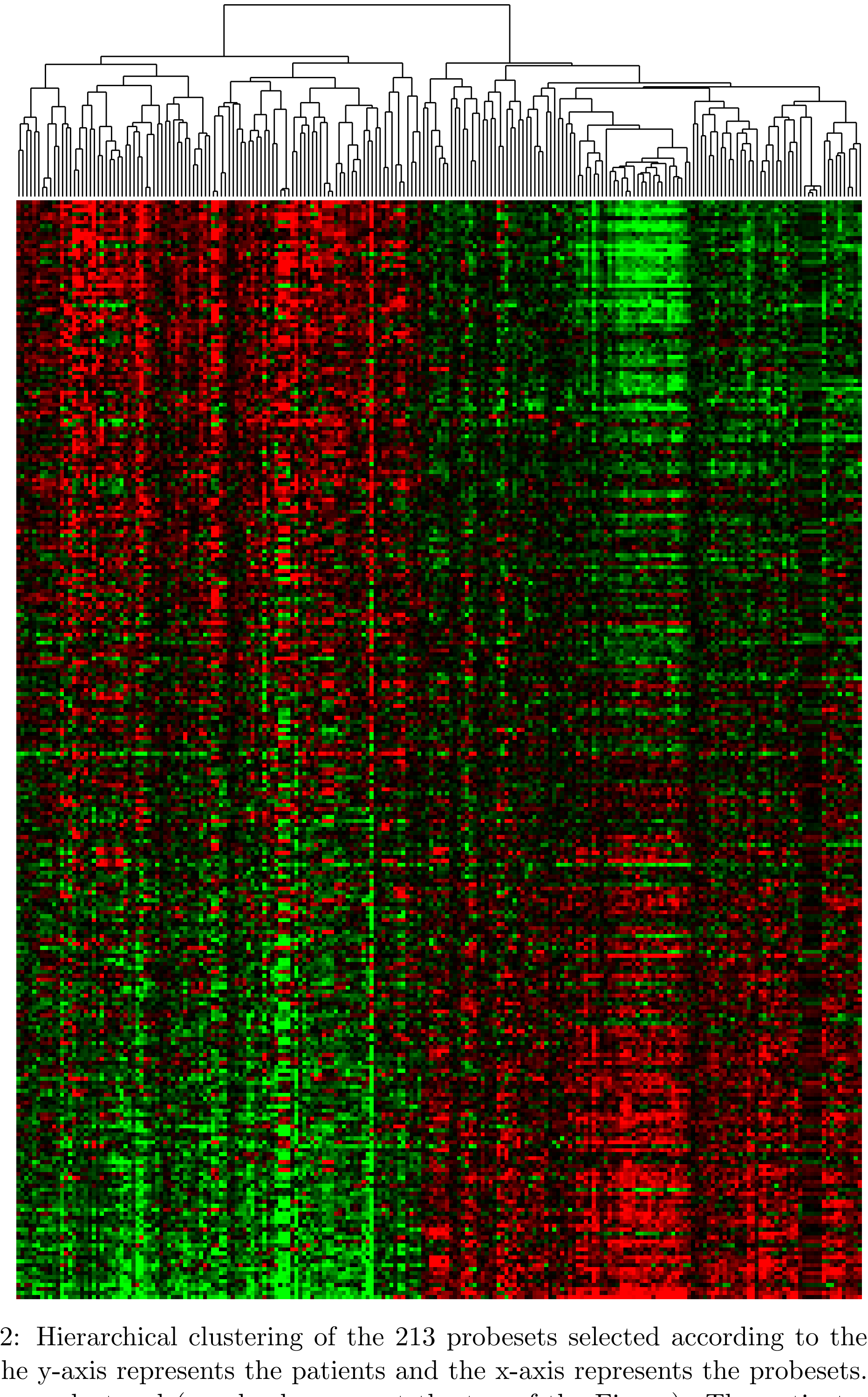


Figure 5.2: Hierarchical clustering of the 213 probesets selected according to their ranking scores. The y-axis represents the patients and the x-axis represents the probesets. Only the probesets are clustered (see dendrogram at the top of the Figure). The patients are sorted by their risk score (see Section 5.1.5), the lowest risk being at the top.

In order to take advantage of the feature selection in multiple microarray platforms comparison (see Section 4.3.2.1), we chose a minimum cluster size of 5 probesets. So the number of constructed features is not equal to the number of clusters (see Section 4.3.2). When a large number of clusters is tested (the number of clusters can be as large as the number of probesets), no feature can be constructed because all the clusters are too small. The relation between the number of clusters and the number of constructed features (with the minimum cluster size set to 5) is given in Figure 5.3. If such a parameter is set to zero, there is no constraint and the number of clusters equals to number of constructed feature. In our case, we can see that the number of constructed features increase rapidly with the number of clusters and starts to decrease when the number of clusters is too large (the size of some clusters decreases below the limit).

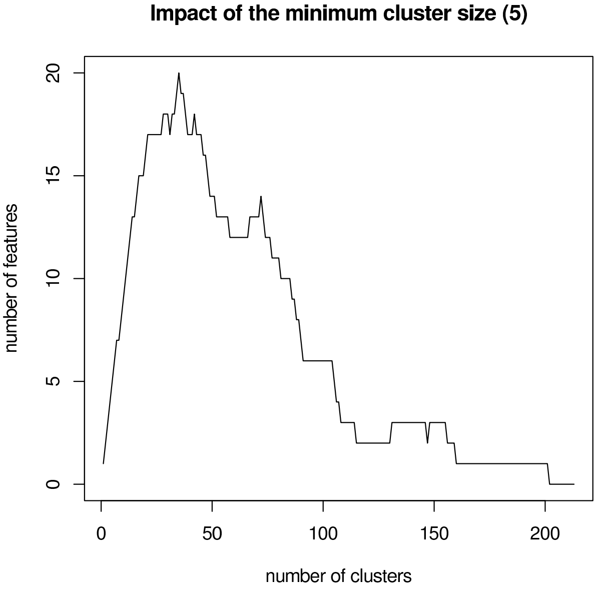


Figure 5.3: Impact of a minimum cluster size set to 5 on the relation between number of clusters and number of constructed features.

The performance of the multivariate Cox model with a set of constructed features, is assessed using 10-fold cross-validation (see Section 4.3.3). The evolution of the training error (minus the (normalized)loglikelihood on the training subset) and the test error (minus the (normalized)loglikelihood on the test subset) are given in Figure 5.5 (see Section 2.4.2.1 for the description of the loglikelihood normalization). We can see that the number of clusters minimizing the test error is two (see Figure 5.4 for the two selected clusters).

Even if the training error decreases to fifty (which is its minimum), we can see that the test error starts to increase from three clusters on. This is an evidence of overfitting, the number of constructed features increasing with the number of clusters between zero and fifty (see Figure 5.3).

5.1.5 Final Cox Model and Risk Score Computation

Once the best set of features is selected (here we have only two features), such features are constructed and used to fit a Cox model on all the training set (see Section 4.4.1).

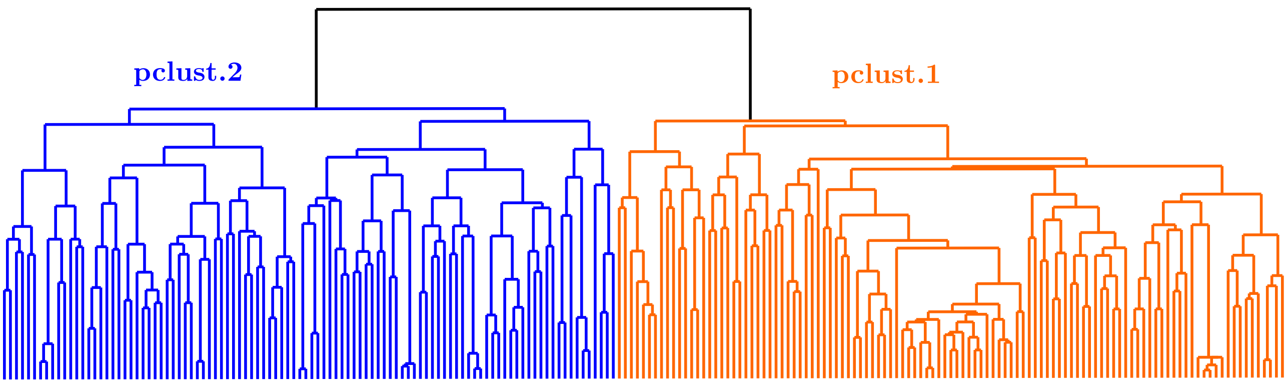


Figure 5.4: Dendrogram with the two selected clusters in color (orange for the cluster 1 and blue for the cluster 2).

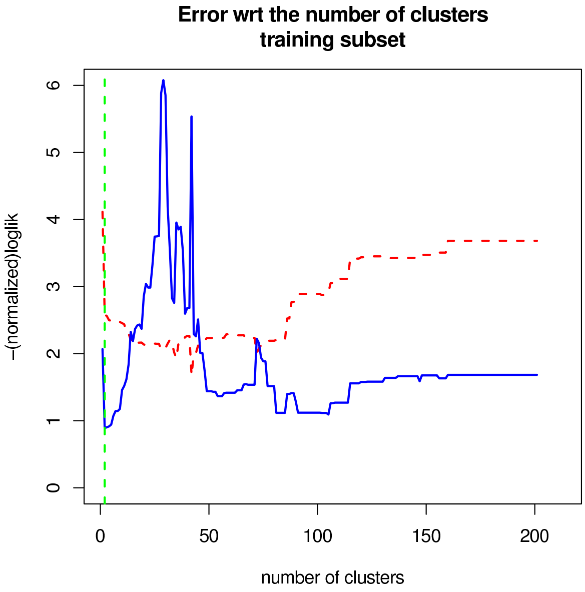


Figure 5.5: Evolution of the error with the number of clusters. The dashed line represents the training error and the solid line represents the test error. The vertical dashed line is the best number of clusters (2) w.r.t. the test error.

```
           coef exp(coef) se(coef)     z       p
pclust.1  1.53   4.59802     1.36  1.12 0.26000
pclust.2 -5.03   0.00653     1.50 -3.34 0.00082

Likelihood ratio test= 59.9  on 2 df,   p=9.77e-14
```

Note that the coefficient of the feature constructed from all the probesets of the cluster 2 (called *pclust.2*), when adjusted with the other feature, is highly significant[2] (p-value = 8.2e-4) while this is not the case for the coefficient of cluster 1 (called *pclust.1*). However, according to the feature selection step (see Section 4.3) and the method used to construct features (see Section 4.3.2), this couple of features is the best one. Moreover the model is significantly different from a model without any features (p-value = 9.77e-14) according to the likelihood ratio test.

Using this model, a risk score is computed for each patient of the training set (see Section 4.4.1). An histogram of risk scores for these patients is given in Figure 5.6. We can see that the approximate distribution given by the histogram is left-skewed, meaning that there are more patients with low risk (in agreement with current clinical observations).

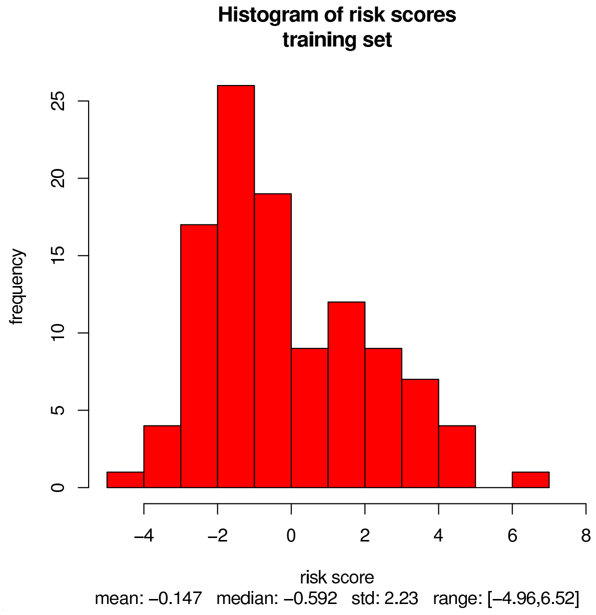


Figure 5.6: Histogram of risk scores for all the patients in the training set. Common statistics are displayed below the histogram.

5.1.6 Cutoff Selection

The selection of a cutoff on the training set is based on the hazard ratio. We compute the other survival statistics described in Section 4.5 in order to assess the effectiveness of the selection procedure.

[2] The p-value is computed by the Wald test from the z statistic and a χ^2 distribution. The likelihood ratio test and the Wald test are described in Section 2.4.2.2.

5.1.6.1 Hazard Ratio

For each possible cutoff, we compute the hazard ratio (see Section 4.5.1). The evolution of the hazard ratio w.r.t. the cutoffs is given in Figure 5.7. The vertical dashed line represents the hazard ratio based cutoff (0.93) selected by the algorithm 3 described in section 4.4.2. The hazard ratio based cutoff gives an hazard ratio of 60.42 with the 95% confidence interval $[7.99, 456.59]$ on the training set.

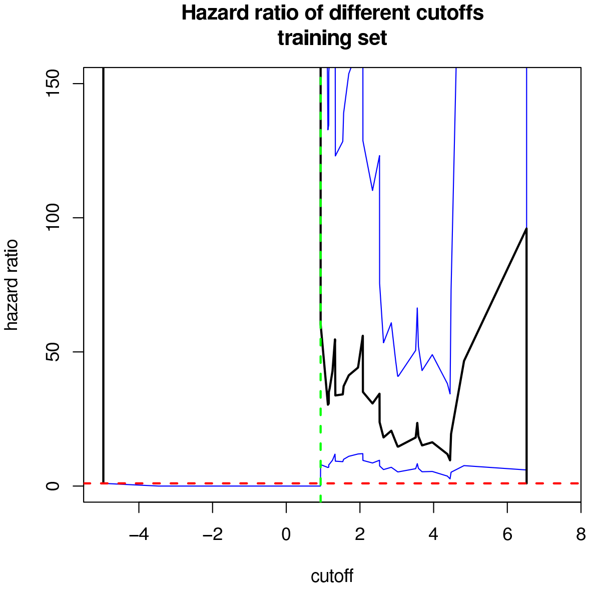


Figure 5.7: Evolution of the hazard ratio w.r.t. the cutoff. The fat line is the HR. The two thin lines delimit the 95% confidence interval around the HR. The horizontal dashed line represents $HR = 1$ which means no difference in survival between low and high-risk groups. The vertical dashed line represents the hazard ratio based cutoff selected by the algorithm 3 in section 4.4.2.

5.1.6.2 Logrank Test

The Figure 5.8 depicts the evolution of the $\log_{10}$ p-value w.r.t. the cutoffs. We can see that the hazard ratio based cutoff gives two groups significantly different according to the logrank test (p-value = 8.37e-13). See Section 4.5.2 for details about the logrank test.

The gaps of the line in Figure 5.8 are due to null p-values, giving $-\inf$ in $\log_{10}$. Such p-values are not plotted.

5.1.6.3 Proportion of DMFS

The Figure 5.9 depicts the evolution of the DMFS proportion (see Section 4.5.3) in low and high-risk groups w.r.t. the cutoffs. We can see that there is no event in the low-risk group, whereas there is 25% of events before three years in the high-risk group. It is interesting to mention that the two functions are quasi monotonically increasing meaning that the patients are very well classified.

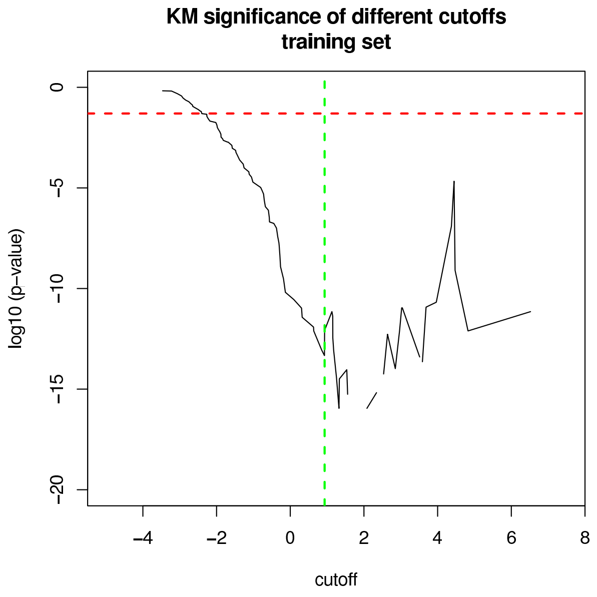


Figure 5.8: Evolution of the $\log_{10}$ p-value w.r.t. the cutoffs. The horizontal dashed line represents the minimum level of significance ($\log_{10}0.05$). The vertical dashed line represents the hazard ratio based cutoff (0.93).

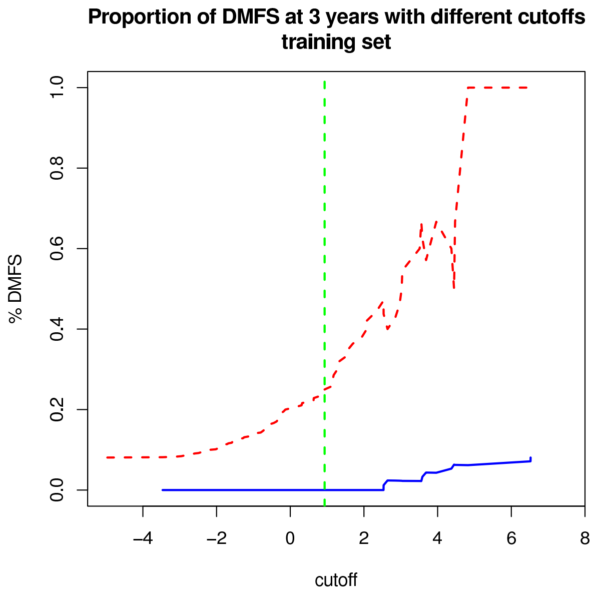


Figure 5.9: Evolution of the DMFS proportion in low (solid line) and high-risk (dashed line) groups w.r.t. the cutoffs. The vertical dashed line represents the hazard ratio based cutoff (0.93).

5.1.6.4 Time-Dependent ROC Curve

The Figure 5.10 gives the time-dependent ROC curve at three years (see Section 4.5.4) in order to assess the performance of the classifier whatever the selected cutoff. The AUC of the classifier equals 0.96 and is different from a random classifier (whose the AUC equals 0.5) represented by the diagonal[3] and shows very good classification performances for events before three years on the training set.

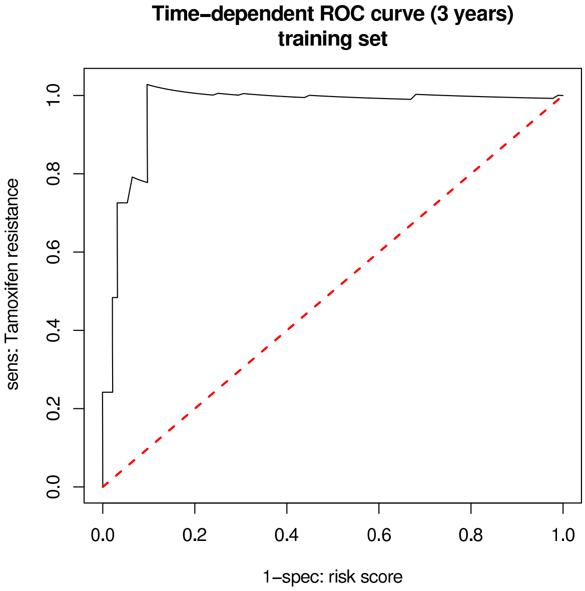


Figure 5.10: Time-dependent ROC curve at three years on the training set. The diagonal (dashed line) represents a random classifier (AUC = 0.5).

Moreover, the AUC is computed for each point in time to highlight the performance of the classifier w.r.t. time (see Figure 5.11). We can see that whatever the point in time, the classifier shows very good performances with AUC > 0.95, especially at three years.

5.1.7 Validation on Independent Test Set

We compute each survival statistic as in the previous section in using the independent test set (patients from KIT and GUYT). The cutoff tested in the next sections is the same than the one selected in the training set.

5.1.7.1 Risk Scores

Using the final model fitted in Section 5.1.5, we compute a risk score for each patient of the test set (see Section 4.4.1). An histogram of risk scores for these patients is given in Figure 5.12. We can see that the approximate distribution given by the histogram is right-skewed, meaning that there are more patients with high risk. It is not the case for the risk scores of the

[3] A p-value addressing the null hypothesis H_o that the area under the ROC curve of the classifier is 0.5 i.e. the AUC of a random classifier, can be computed using the U statistic for Mann-Whitney test [Mason and Graham, 1982].

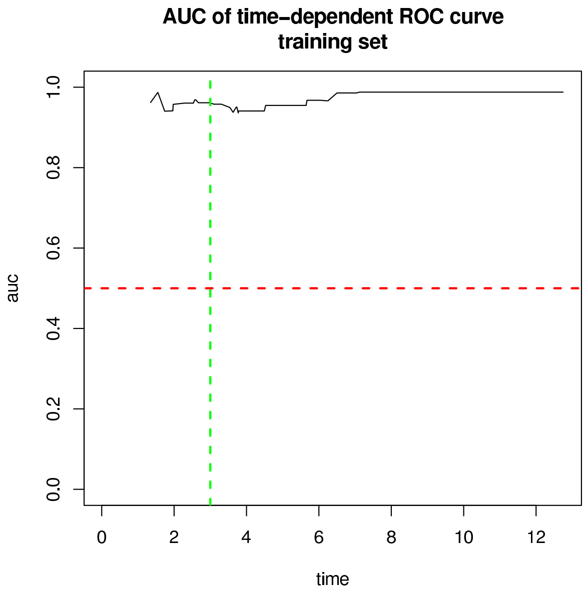


Figure 5.11: Evolution of the AUC's w.r.t. time on the training set. The horizontal dashed line represents the AUC of a random classifier (AUC = 0.5). The vertical dashed line represents the 3 years mark.

patients in the training set (see Section 5.1.5) where we find the contrary. This is an evidence that the training set (OXFT population) and the test set (KIT and GUYT populations) are slightly different[4]. This can lead to a poor validation performances due to differences between training and test sets.

5.1.7.2 Hazard Ratio

For each cutoff tested in the training set, the hazard ratio is computed on the test set. The evolution of the hazard ratio w.r.t. the cutoffs is given in Figure 5.13. The vertical dashed line represents the hazard ratio based cutoff (0.93) selected previously. This cutoff gives an hazard ratio of 2.44 with the 95% confidence interval $[1.38, 4.31]$ on the test set. The 95% confidence interval does not include the unity.

5.1.7.3 Logrank Test

The Figure 5.14 depicts the evolution of the $\log_{10}$ p-value w.r.t. the cutoffs on the test set. We can see that the hazard ratio based cutoff gives two groups significantly different according to the logrank test (p-value = 1.47e-3). See Section 4.5.2 for details about the logrank test.

5.1.7.4 Proportion of DMFS

The Figure 5.15 depicts the evolution of the DMFS proportion in low and high-risk groups w.r.t. the cutoffs on the test set. We can see that there is 8% of events in the low-risk group,

[4] A careful study of the demographic data shows that the KIT population contains a lot of early distant metastases (early events) in comparison to the OXFT and GUYT populations.

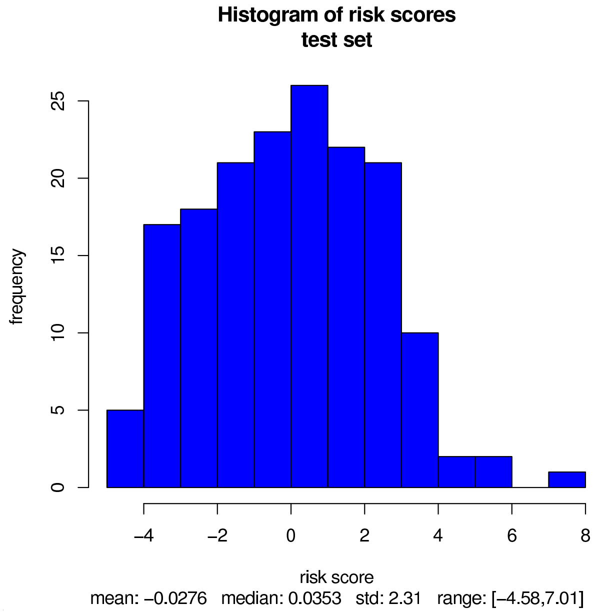


Figure 5.12: Histogram of risk scores for all the patients in the test set. Common statistics are displayed below the histogram.

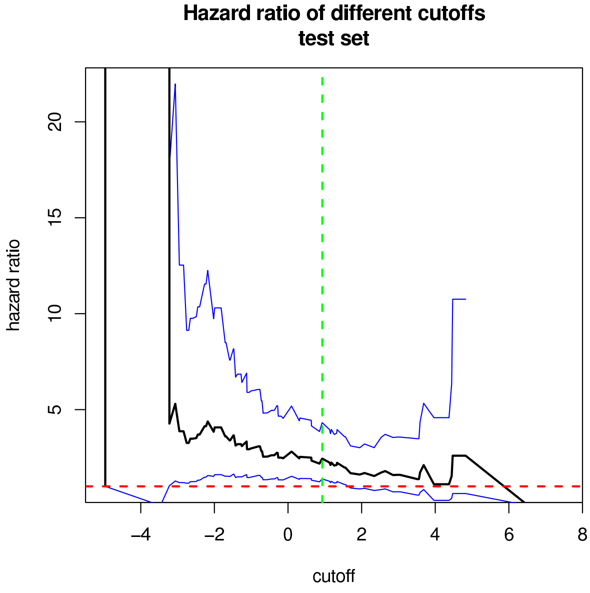


Figure 5.13: Evolution of the hazard ratio w.r.t. the cutoffs. The fat line is the HR. the two thin lines is the 95% confidence interval around the HR. The horizontal dashed line represents $HR = 1$ which means no difference in survival between low and high-risk groups. The vertical dashed line represents the hazard ratio based cutoff selected by the algorithm 3 in Section 4.4.2.

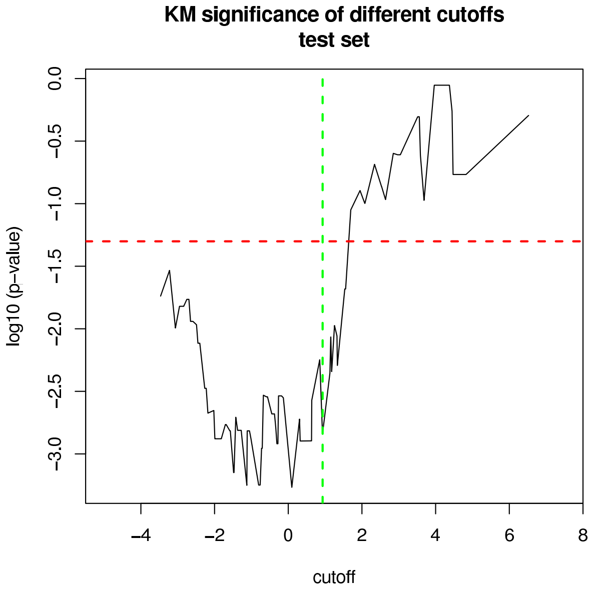


Figure 5.14: Evolution of the $\log_{10}$ p-value w.r.t. the cutoffs. The horizontal dashed line represents the minimum level of significance ($\log_{10}0.05$). The vertical dashed line represents the hazard ratio based cutoff (0.93).

whereas there is 20% of event in the high-risk group, before three years. The very high-risk patients are not well classified as the Figure 5.15 shows (if we choose a cutoff of 4, there is no more event before three years in the high-risk group).

5.1.7.5 Time-Dependent ROC Curve

The time-dependent ROC curve at three years on the test set is given in Figure 5.16. The AUC of the classifier equals to 0.65, remaining different than a random one but the difference is less evident.

The Figure 5.17 depicts the evolution of the AUC w.r.t. time on the test set. Interestingly, the classifier has poor performances on very early events (in the first year) but gives much better performances after three years. This fact need to be investigated in further analysis.

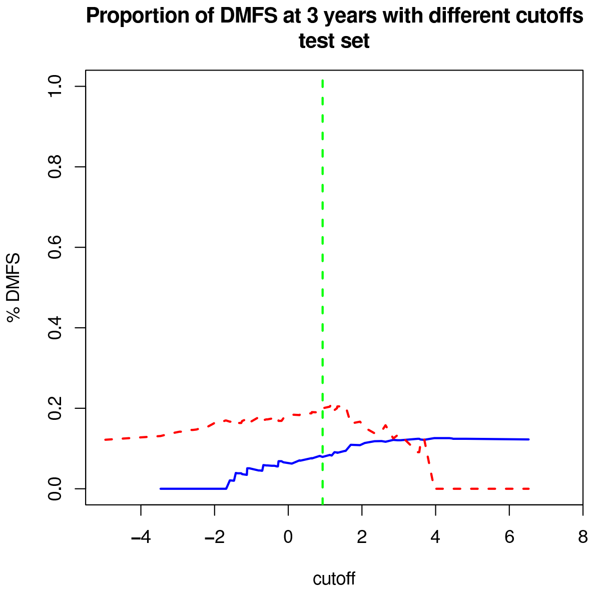


Figure 5.15: Evolution of the DMFS proportion in low (solid line) and high-risk (dashed line) groups w.r.t. the cutoffs. The vertical dashed line represents the hazard ratio based cutoff (0.93).

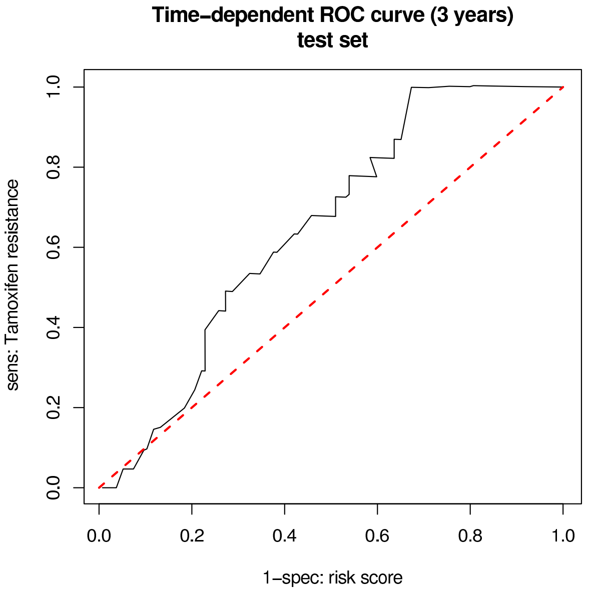


Figure 5.16: Time-dependent ROC curve at three years on the test set. The diagonal (dashed line) represents a random classifier (AUC = 0.5).

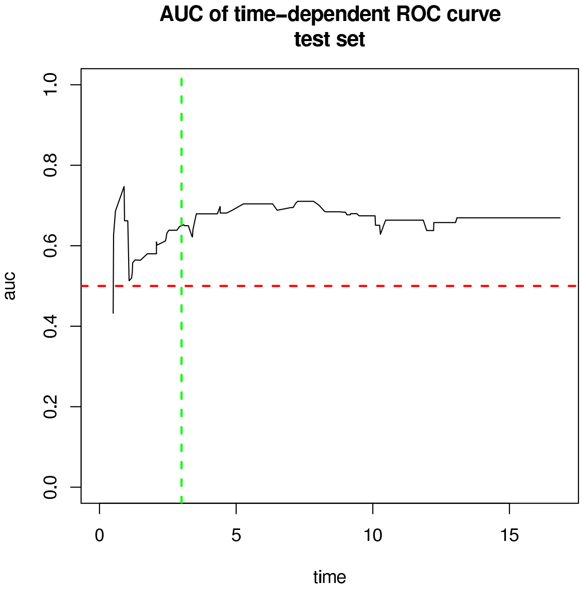


Figure 5.17: Evolution of the AUC's w.r.t. time on the test set. The horizontal dashed line represents the AUC of a random classifier (AUC = 0.5). The vertical dashed line represents the 3 years mark.

Chapter 6

Conclusion

Contents

We have proposed a machine learning methodology for the microarray analysis using machine learning and well-established survival methods. This methodology covers the whole range of such an analysis, starting from the raw data and their preprocessing to end with high-level analysis like the feature selection, the construction of the classifier on the training set and its validation on independent test set using several traditional survival statistics.

We have chosen to use a classifier based on survival analysis instead of a binary classifier as in [van't Veer et al., 2002] for instance. Indeed, when we transform the survival data in binary classes, we loose information (see Chapter 2). We prefer to use all the information available in the survival data.

Moreover, we have chosen to develop a methodology keeping the classifier interpretable instead of using it as a black box. From a biological point of view, we can extract interesting biological information from the final classifier. Indeed the risk computation is based on a linear combination of several weighted cluster centroids (see Section 4.3.2). The cluster centroids are the average of several probesets and the weights are the coefficients fitted by the Cox regression. So we can study in details the biological information of such clusters and their contributions in the risk computation.

We have tested the methodology on real data dealing with the Tamoxifen resistance of breast cancer patients. We have constructed a classifier based on a training set of 99 patients, able to assess correctly the risk of the patients in the test set (156 patients), and then classify them in low and high-risk groups. This classifier results in a hazard ratio of 2.44 with the 95% confidence interval $[1.38, 4.31]$ between the two groups (see Figure 5.13). This difference in survival is confirmed by the logrank test (p-value = 1.47e-3, see Figure 5.14). Moreover, there is a very low percentage of distant metastases in the low-risk group within the first three years (8% in 101 patients) whereas there is 20% (in 55 patients) of distant metastases before three years in the high-risk group (see Figure 5.15). The evolution of the AUC of time-dependent ROC curves shows us that the classifier has poor performance for very early distant metastases (within the first year) but has good performance for later events (including the three years mark, see Figure 5.17).

Even if the results seem to be very promising, there exist numerous alternatives in terms of methods used for the variable ranking, the feature construction and the classifier. We can test them and compare the different results. Moreover, we can test the classifier on different datasets if publicly available and assess the benefits of the feature construction described in Section 4.3.2.1.

6.1 Future Works

1. Study of the impact of the normalization methods on the results. Currently, new normalization methods are introduced (e.g. GCRMA[1]) and recent articles challenge the performance assessment of such methods [Ploner et al., 2005; Sheden et al., 2005; Gautier et al., 2005].

2. Implementation of data preprocessing methods for specific computer architectures like computers clusters. Indeed, the data preprocessing step is computational intensive and current tools are inefficient for larger datasets. These tools need to be adapted to specific computer architecture.

3. Study of the variance of variable ranking methods and the overlap of the results between different populations of patients. Variable ranking is commonly used in microarray studies without any consideration of their intrinsic variance within one population and inter-population. The microarray data available at the Microarray Unit (IJB) gives us the opportunity to carry out such an analysis.

4. Use of alternative methods for the feature construction :

   (a) Use of alternative clustering methods like the adaptive quality-based clustering [De Smet et al., 2002] instead of hierarchical clustering. Such a method has not the constraint of clustering all the variables. So, only highly correlated variables will be clustered together whereas uncorrelated variables will be left alone.

   (b) Use of methods of space dimensionality reduction and input variable transformation (e.g. PCA) to construct new features which are independent of each others.

5. Study of the classifier robustness with the loss of one or more probes. As described in Section 4.3.2.1, we may lose some variables used to construct features when we test the classifier on another microarray platform. It would be interesting to study the performance of the classifier in removing one or more probesets to assess its robustness.

6. Use of penalized Cox model [Tibshirani, 1997; Gui and Li, 2004] in order to perform a feature selection at the level of the model (see embedded methods described in Section 2.6.2.1).

7. Currently, we have studied the performance of the survival analysis design on one split of the data based on the populations. Indeed, the training and the test sets are composed of one or more populations without mixing the patients between populations. A multiple random validation strategy [Michiels et al., 2005] is necessary to assess the performance whatever the training/test split.

[1] GCRMA is a software package introduced by W. Zhijin in 2003 (`http://www.bioconductor.org/repository/release1.4/package/html/gcrma.html`).

8. Comparison with binary classification techniques [Brown et al., 1999; Duda et al., 2001; Dudoit et al., 2002].

9. Use of Gene Ontology (see Appendix D) to infer biological knowledge about the probesets selected to construct the features.

10. Comparison with traditional histological criteria and consensus (see Chapter 1).

11. Comparison with other molecular signatures using different technologies [Paik et al., 2004; Ma et al., 2004].

12. Test of the developed classifier on other datasets if publicly available.

Appendix A

Semiparametric Regression Models : Additional Topics

In this chapter, a number of additional topics that arise in the practical application of the semiparametric regression models are discussed.

A.1 Tied Data

The formula for the partial likelihood in (2.21) is valid only for data in which no two events occur at the same time. However, it is quite common for data to contain tied event times, so an alternative formula is needed to handle those situations. The most common alternative is called the *Breslow's approximation*, which works well with relatively few ties. When data are heavily tied, the approximation can be quite poor [Farewell and Prentice, 1980; Hsieh, 1995]. There exist better approximations proposed by [Efron, 1977] as well as the *exact* and the *discrete* methods. More details are given in [Therneau and Grambsch, 2000].

A.2 Time-Dependent Covariate

The time-dependent covariates may change in value over the time of observation. While it is simple to modify Cox's model to allow for time-dependent covariates, the computations of the resulting partial likelihood is much more time consuming.

To modify the model in (2.15) to include time-dependent covariates, all we need to do is write (t) after the x's that are time dependent. For a model with one fixed covariate and one time-dependent covariate, we have

$$\ln h_i(t) = \alpha(t) + \beta_1 x_{i1} + \beta_2 x_{i2}(t) \tag{A.1}$$

The hazard at time t depends on the value of x_1, and on the value of x_2 at time t. $x_2(t)$ can be defined using any information about the individual prior to time t. The computation of each time-dependent covariate at each time t can be expensive.

A.3 Nonproportional Hazards

When time-dependent covariates are introduced in the Cox regression model, the assumption of proportional hazards is violated. Indeed, because the time-dependent covariates will change at different rates for different individuals, the ratios of their hazards can not remain constant. However, we have seen in Section A.2 that the partial likelihood can treat such situations.

Proportional hazards assumption violations for fixed covariates are equivalent to interactions between one or more covariates and time. The proportional hazards model assumes that the effect of each covariate is the same at all points in time. If the effect of a covariate varies with time, the proportional hazards assumption is violated for that variable.

Explicit Interaction Method A common way of representing interaction between two variables in a linear regression model is to include a new variable that is the product of the two variables in equation (see Section A.2). To represent the interaction between a covariate x and time in a Cox model, we can write

$$\ln h(t) = \alpha(t) + \beta_1 x + \beta_2 xt \tag{A.2}$$

Factoring out the x , we can write this as

$$\ln h(t) = \alpha(t) + (\beta_1 + \beta_2 t)\, x \tag{A.3}$$

In this equation the effect of x is $\beta_1 + \beta_2 t$. If β_2 is positive, then the effect of x increase linearly with time; if it is negative, the effect decreases with time linearly with time. β_1 can be interpreted as the effect of x at time 0, the origin of the process. This model can be easily estimated by defining a time-dependent covariate $z = xt$. It is also straightforward to include interactions between time and time-dependent covariates. Then, the covariates will change over time but also the effect of those covariates will change over time.

In order to test the proportional hazards assumption, a time-dependent covariate representing the interaction of the original covariate and time, can be added to the model for any suspected covariate. If the interaction covariate have a significant coefficient, we have evidence for nonproportionality. Otherwise, we may conclude than the proportional hazards assumption is not violated.

Stratification Method Another approach to nonproportionality is *stratification*, a technique that is most useful when the covariate that interacts with time is both categorical and not of direct interest. Let z be such a binary covariate and we suspect that the effect of z varies with time. Alternatively, we can say that the shape of the hazard function is different according to z. Let x be another covariate of the model. The model can be written as

$$\begin{cases} \ln h_i(t) = \alpha_0(t) + \beta x_i \text{ if } z = 0 \\ \ln h_i(t) = \alpha_1(t) + \beta x_i \text{ if } z = 1 \end{cases}$$

Notice that the coefficient of x is the same in both equations, but the arbitrary function of time is allowed to to differ. We can combine the two equations into a single equation by writing

$$\ln h_i(t) = \alpha_z(t) + \beta x_i$$

The model can be estimated by the method of partial likelihood using these steps

1. Construct separate partial likelihood functions for each of the value of z.

2. Multiply those two functions together.

3. Choose values of β that maximize this function.

If the coefficient of the covariate x is not significant in the model including x and z but is in the model including x and stratified by z, we can conclude that it is important to control the effect of the covariate z.

Compared to the explicit interaction method, the method of stratification has two main advantages :

- The explicit interaction method requires to choose a particular form for the interaction, but stratification allows for any kind of change in the effect of a covariate over time.
- Stratification is easier to set up ans is less expensive in computation time.

But there are also important disadvantages of the stratification :

- There is no way to test for either the main effect of the stratifying covariate or its interaction with time. In particular, it is not legitimate to compare the log-likelihoods for models with and without a stratifying covariate [Allison, 1995].
- If the form of the interaction with time is correctly specified, the explicit interaction method should yield more efficient estimates of the coefficients of the other covariates.

A.4 Estimating Survivor Functions

The form of the dependence of the hazard on time is left unspecified in the proportional hazards model. Furthermore, the partial likelihood method discards that portion of the likelihood that contains information about the dependence of the hazard on time. Nevertheless, it is possible to get nonparametric estimates of the survivor function based on a fitted proportional hazards model.

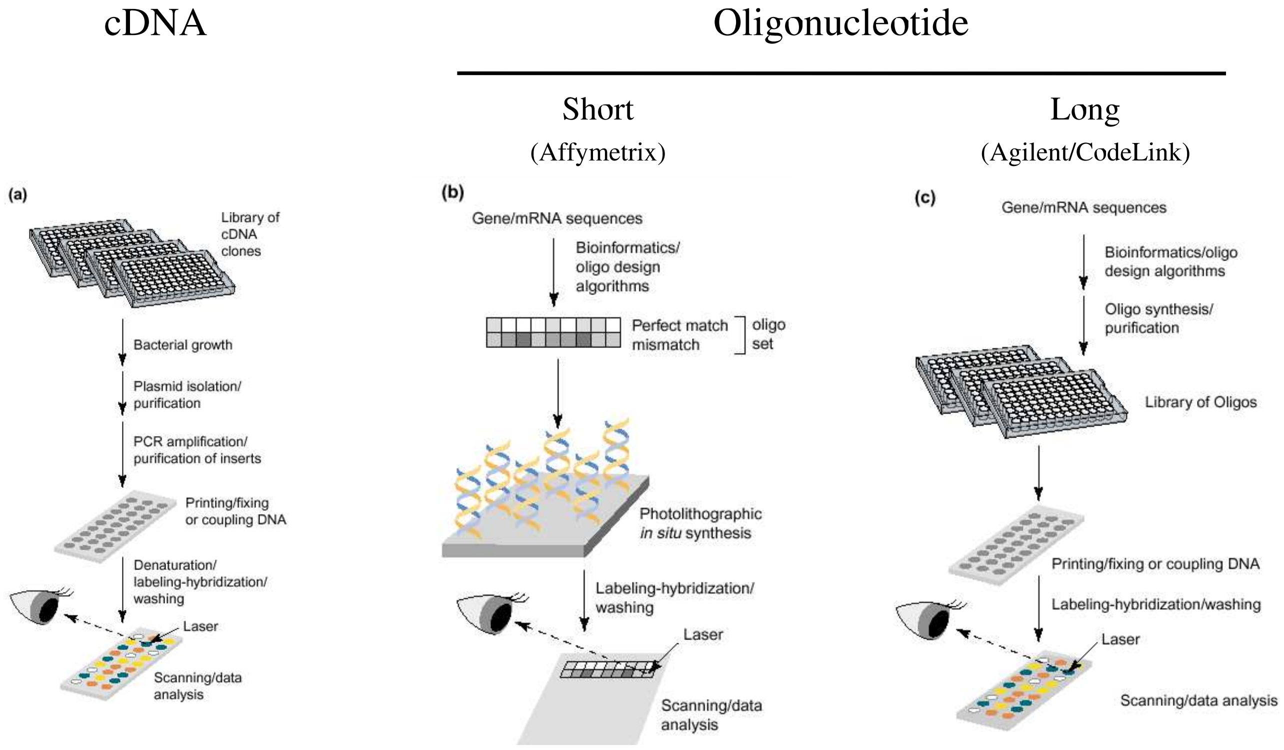
When there are no time-dependent covariates, the Cox model can be written as

$$S(t) = [S_0(t)]^{\exp(\boldsymbol{\beta}\mathbf{x})}$$

where $S(t)$ is the survival probability at time t for an individual with covariate values $\mathbf{x}$, and $S_0(t)$ is the *baseline survivor function*, that is the survivor function for an individual whose covariate values are all zero. After estimating $\boldsymbol{\beta}$ by partial likelihood, we can get an estimate of $S_0(t)$ by a nonparametric maximum likelihood method (see [Collett, 2003] for details).

Appendix B

Microarray Platforms


cDNA
Oligonucleotide
Short
(Affymetrix)
Long
(Agilent/CodeLink)
(a)
Library of cDNA clones
Bacterial growth
Plasmid isolation/ purification
PCR amplification/ purification of inserts
Printing/fixing or coupling DNA
Denaturation/ labeling-hybridization/ washing
Laser
Scanning/data analysis
(b)
Gene/mRNA sequences
Bioinformatics/ oligo design algorithms
Perfect match
mismatch
oligo set
Photolithographic *in situ* synthesis
Labeling-hybridization/ washing
Laser
Scanning/data analysis
(c)
Gene/mRNA sequences
Bioinformatics/oligo design algorithms
Oligo synthesis/ purification
Library of Oligos
Printing/fixing or coupling DNA
Labeling-hybridization/washing
Laser
Scanning/data analysis

Overview of different microarray technologies.

Appendix C

Probeset Annotations

probeset	accession number	gene name	symbol	unigene	pclust
227578_at	H28597	thymopoietin	TMPO	Hs.11355	pclust.1
201014_s_at	NM_006452	phosphoribosylaminoimidazole carboxylase, phosphoribosylaminoimidazole succinocarboxamide synthetase	PAICS	Hs.518774	pclust.1
225723_at	BE794699	chromosome 6 open reading frame 129	C6orf129	Hs.284207	pclust.1
204033_at	NM_004237	thyroid hormone receptor interactor 13	TRIP13	Hs.436187	pclust.1
208696_at	AF275798	chaperonin containing TCP1, subunit 5 (epsilon)	CCT5	Hs.1600	pclust.1
200750_s_at	AF054183	RAN, member RAS oncogene family	RAN	Hs.519656	pclust.1
213911_s_at	BF718636	H2A histone family, member Z	H2AFZ	Hs.119192	pclust.1
204331_s_at	NM_021107	mitochondrial ribosomal protein S12	MRPS12	Hs.411125	pclust.1
202433_at	NM_005827	solute carrier family 35, member B1	SLC35B1	Hs.154073	pclust.1
226943_at	AA287457	NA	NA	NA	pclust.1
202779_s_at	NM_014501	ubiquitin-conjugating enzyme E2S	UBE2S	Hs.396393	pclust.1
201947_s_at	NM_006431	chaperonin containing TCP1, subunit 2 (beta)	CCT2	Hs.189772	pclust.1
200853_at	NM_002106	H2A histone family, member Z	H2AFZ	Hs.119192	pclust.1
209714_s_at	AF213033	cyclin-dependent kinase inhibitor 3 (CDK2-associated dual specificity phosphatase)	CDKN3	Hs.84113	pclust.1
201764_at	NM_024056	hypothetical protein MGC5576	MGC5576	Hs.103834	pclust.1
201475_x_at	NM_004990	methionine-tRNA synthetase	MARS	Hs.355867	pclust.1
222077_s_at	AU153848	Rac GTPase activating protein 1	RACGAP1	Hs.505469	pclust.1
219588_s_at	NM_017760	more than blood homolog	MTB	Hs.18616	pclust.1
238728_at	AA194266	NA	NA	NA	pclust.1

probeset	accession number	gene name	symbol	unigene	pclust
201342_at	NM_003093	small nuclear ribonucleoprotein polypeptide C	SNRPC	Hs.1063	pclust.1
204962_s_at	NM_001809	centromere protein A, 17kDa	CENPA	Hs.1594	pclust.1
218009_s_at	NM_003981	protein regulator of cytokinesis 1	PRC1	Hs.459362	pclust.1
234347_s_at	AF038554	density-regulated protein	DENR	Hs.22393	pclust.1
201606_s_at	BE796924	nuclear phosphoprotein similar to S. cerevisiae PWP1	PWP1	Hs.506652	pclust.1
219060_at	NM_018024	hypothetical protein FLJ10204	FLJ10204	Hs.18029	pclust.1
203011_at	NM_005536	inositol(myo)-1(or 4)-monophosphatase 1	IMPA1	Hs.492120	pclust.1
219275_at	NM_004708	programmed cell death 5	PDCD5	Hs.482549	pclust.1
204092_s_at	NM_003600	aurora kinase A	AURKA	NA	pclust.1
205046_at	NM_001813	centromere protein E, 312kDa	CENPE	Hs.75573	pclust.1
228357_at	BE966979	NA	NA	NA	pclust.1
226287_at	AI458313	hypothetical protein AF301222	LOC81023	Hs.143733	pclust.1
235704_at	AI307251	DAZ associated protein 2	DAZAP2	Hs.369761	pclust.1
218782_s_at	NM_014109	ATPase family, AAA domain containing 2	ATAD2	Hs.370834	pclust.1
239753_at	BE560888	NA	NA	NA	pclust.1
217932_at	NM_015971	mitochondrial ribosomal protein S7	MRPS7	Hs.71787	pclust.1
213008_at	BG403615	hypothetical protein FLJ10719	FLJ10719	Hs.513126	pclust.1
211058_x_at	BC006379	tubulin, alpha, ubiquitous	K-ALPHA-1	Hs.524390	pclust.1
224913_s_at	AA877820	translocase of inner mitochondrial membrane 50 homolog (yeast)	TIMM50	Hs.355819	pclust.1
214710_s_at	BE407516	cyclin B1	CCNB1	Hs.23960	pclust.1
218883_s_at	NM_024629	MLF1 interacting protein	MLF1IP	Hs.481307	pclust.1
223785_at	BC004277	hypothetical protein FLJ10719	FLJ10719	Hs.513126	pclust.1
218556_at	NM_014182	ORM1-like 2 (S. cerevisiae)	ORMDL2	Hs.534450	pclust.1
205543_at	NM_014278	heat shock 70kDa protein 4-like	HSPA4L	Hs.135554	pclust.1
210334_x_at	AB028869	baculoviral IAP repeat-containing 5 (survivin)	BIRC5	Hs.514527	pclust.1
208079_s_at	NM_003158	serine/threonine kinase 6	STK6	Hs.250822	pclust.1
218027_at	NM_014175	mitochondrial ribosomal protein L15	MRPL15	Hs.18349	pclust.1
211072_x_at	BC006481	tubulin, alpha, ubiquitous	K-ALPHA-1	Hs.524390	pclust.1
209251_x_at	BC004949	tubulin alpha 6	TUBA6	Hs.436035	pclust.1
204825_at	NM_014791	maternal embryonic leucine zipper kinase	MELK	Hs.184339	pclust.1
210821_x_at	BC002703	centromere protein A, 17kDa	CENPA	Hs.1594	pclust.1
226604_at	AA418403	NA	NA	NA	pclust.1
209218_at	AF098865	squalene epoxidase	SQLE	Hs.71465	pclust.1
219661_at	NM_022897	RAN binding protein 17	RANBP17	Hs.410810	pclust.1
225702_at	AA973041	hypothetical protein FLJ14825	FLJ14825	Hs.521800	pclust.1

probeset	accession number	gene name	symbol	unigene	pclust
222992_s_at	AF261090	NADH dehydrogenase (ubiquinone) 1 beta subcomplex, 9, 22kDa	NDUFB9	Hs.15977	pclust.1
201292_at	AL561834	topoisomerase (DNA) II alpha 170kDa	TOP2A	Hs.156346	pclust.1
213310_at	AI613483	eukaryotic translation initiation factor 2C, 2	EIF2C2	Hs.449415	pclust.1
201090_x_at	NM_006082	tubulin, alpha, ubiquitous	K-ALPHA-1	Hs.524390	pclust.1
212722_s_at	AK021780	phosphatidylserine receptor	PTDSR	Hs.514505	pclust.1
208838_at	AB020636	TBP-interacting protein	TIP120A	Hs.546407	pclust.1
203554_x_at	NM_004219	pituitary tumor-transforming 1	PTTG1	Hs.350966	pclust.1
220060_s_at	NM_017915	hypothetical protein FLJ20641	FLJ20641	Hs.330663	pclust.1
212021_s_at	AU132185	antigen identified by monoclonal antibody Ki-67	MKI67	Hs.80976	pclust.1
218326_s_at	NM_018490	leucine-rich repeat-containing G protein-coupled receptor 4	LGR4	Hs.502176	pclust.1
224330_s_at	AB049647	mitochondrial ribosomal protein L27	MRPL27	Hs.7736	pclust.1
216125_s_at	AF064606	RAN binding protein 9	RANBP9	Hs.306242	pclust.1
203880_at	NM_005694	COX17 homolog, cytochrome c oxidase assembly protein (yeast)	COX17	Hs.534383	pclust.1
226376_at	AI885018	zinc finger CCCH type domain containing 5	ZC3HDC5	Hs.201859	pclust.1
212500_at	AL049319	chromosome 10 open reading frame 22	C10orf22	Hs.99821	pclust.1
203764_at	NM_014750	discs, large homolog 7 (Drosophila)	DLG7	Hs.77695	pclust.1
203007_x_at	AF077198	lysophospholipase I	LYPLA1	Hs.435850	pclust.1
206698_at	NM_021083	Kell blood group precursor (McLeod phenotype)	XK	Hs.78919	pclust.1
223110_at	BC003701	DKFZP434I116 protein	DKFZP434I1	Hs.202238	pclust.1
224331_s_at	AB049654	mitochondrial ribosomal protein L36	MRPL36	Hs.32196	pclust.1
203214_x_at	NM_001786	cell division cycle 2, G1 to S and G2 to M	CDC2	Hs.334562	pclust.1
213741_s_at	BF575685	karyopherin alpha 1 (importin alpha 5)	KPNA1	Hs.161008	pclust.1
218046_s_at	NM_016065	mitochondrial ribosomal protein S16	MRPS16	Hs.180312	pclust.1
217946_s_at	NM_016402	SUMO-1 activating enzyme subunit 1	SAE1	Hs.515500	pclust.1
200659_s_at	NM_002634	prohibitin	PHB	Hs.514303	pclust.1
200925_at	NM_004373	cytochrome c oxidase subunit VIa polypeptide 1	COX6A1	Hs.497118	pclust.1
223156_at	BC000242	mitochondrial ribosomal protein S23	MRPS23	Hs.5836	pclust.1

probeset	accession number	gene name	symbol	unigene	pclust
215452_x_at	AL031133	SMT3 suppressor of mif two 3 homolog 2 (yeast)	SUMO2	Hs.546298	pclust.1
210216_x_at	AF084513	RAD1 homolog (S. pombe)	RAD1	Hs.547084	pclust.1
213379_at	AF091086	hypothetical protein CL640	CL640	Hs.144304	pclust.1
200932_s_at	NM_006400	dynactin 2 (p50)	DCTN2	Hs.289123	pclust.1
234464_s_at	AK021607	essential meiotic endonuclease 1 homolog 1 (S. pombe)	EME1	Hs.514330	pclust.1
209849_s_at	AF029669	RAD51 homolog C (S. cerevisiae)	RAD51C	Hs.412587	pclust.1
231609_at	AW418674	chromosome 10 open reading frame 82	C10orf82	Hs.121347	pclust.1
222267_at	BE619220	hypothetical protein FLJ14803	FLJ14803	Hs.267245	pclust.1
202095_s_at	NM_001168	baculoviral IAP repeat-containing 5 (survivin)	BIRC5	Hs.514527	pclust.1
209408_at	U63743	kinesin family member 2C	KIF2C	Hs.69360	pclust.1
212160_at	AI984005	exportin, tRNA (nuclear export receptor for tRNAs)	XPOT	Hs.85951	pclust.1
225827_at	AI832074	eukaryotic translation initiation factor 2C, 2	EIF2C2	Hs.449415	pclust.1
211662_s_at	L08666	voltage-dependent anion channel 2	VDAC2	Hs.355927	pclust.1
212022_s_at	BF001806	antigen identified by monoclonal antibody Ki-67	MKI67	Hs.80976	pclust.1
213647_at	D42046	DNA2 DNA replication helicase 2-like (yeast)	DNA2L	Hs.532446	pclust.1
212639_x_at	AL581768	tubulin, alpha, ubiquitous	K-ALPHA-1	Hs.524390	pclust.1
202069_s_at	AI826060	isocitrate dehydrogenase 3 (NAD+) alpha	IDH3A	Hs.546262	pclust.1
242218_at	AI201116	peroxisome proliferative activated receptor, delta	PPARD	Hs.485196	pclust.1
201524_x_at	NM_003348	ubiquitin-conjugating enzyme E2N (UBC13 homolog, yeast)	UBE2N	Hs.524630	pclust.1
202704_at	AA675892	transducer of ERBB2, 1	TOB1	Hs.531550	pclust.1
223472_at	AF071594	Wolf-Hirschhorn syndrome candidate 1	WHSC1	Hs.113876	pclust.1
201597_at	NM_001865	cytochrome c oxidase subunit VIIa polypeptide 2 (liver)	COX7A2	Hs.70312	pclust.1
224753_at	BE614410	cell division cycle associated 5	CDCA5	Hs.434886	pclust.1
219555_s_at	NM_018455	uncharacterized bone marrow protein BM039	BM039	Hs.283532	pclust.1
220318_at	NM_017957	epsin 3	EPN3	Hs.165904	pclust.1
229068_at	BF197357	chaperonin containing TCP1, subunit 5 (epsilon)	CCT5	Hs.1600	pclust.1
202954_at	NM_007019	ubiquitin-conjugating enzyme E2C	UBE2C	Hs.93002	pclust.1
201483_s_at	BC002802	suppressor of Ty 4 homolog 1 (S. cerevisiae)	SUPT4H1	Hs.439481	pclust.1

probeset	accession number	gene name	symbol	unigene	pclust
201804_x_at	NM_001281	cytoskeleton associated protein 1	CKAP1	Hs.31053	pclust.1
221520_s_at	BC001651	cell division cycle associated 8	CDCA8	Hs.524571	pclust.1
235427_at	AA418074	NA	NA	NA	pclust.2
208892_s_at	BC003143	dual specificity phosphatase 6	DUSP6	Hs.298654	pclust.2
225197_at	W58461	NA	NA	NA	pclust.2
218983_at	NM_016546	complement component 1, r subcomponent-like	C1RL	Hs.525264	pclust.2
208891_at	BC003143	dual specificity phosphatase 6	DUSP6	Hs.298654	pclust.2
202018_s_at	NM_002343	lactotransferrin	LTF	Hs.529517	pclust.2
213107_at	R59093	TRAF2 and NCK interacting kinase	TNIK	Hs.34024	pclust.2
219281_at	NM_012331	methionine sulfoxide reductase A	MSRA	Hs.490981	pclust.2
204015_s_at	BC002671	dual specificity phosphatase 4	DUSP4	Hs.417962	pclust.2
224835_at	AL109935	ribosomal protein S18 pseudogene 1	RPS18P1	NA	pclust.2
209940_at	AF083068	poly (ADP-ribose) polymerase family, member 3	PARP3	Hs.271742	pclust.2
205898_at	U20350	chemokine (C-X3-C motif) receptor 1	CX3CR1	Hs.78913	pclust.2
214175_x_at	AI254547	PDZ and LIM domain 4	PDLIM4	Hs.424312	pclust.2
202962_at	NM_015254	kinesin family member 13B	KIF13B	Hs.444767	pclust.2
205011_at	NM_014622	loss of heterozygosity, 11, chromosomal region 2, gene A	LOH11CR2A	Hs.152944	pclust.2
226034_at	BE222344	NA	NA	NA	pclust.2
200762_at	NM_001386	dihydropyrimidinase-like 2	DPYSL2	Hs.173381	pclust.2
209295_at	AF016266	tumor necrosis factor receptor superfamily, member 10b	TNFRSF10B	Hs.521456	pclust.2
214486_x_at	AF041459	CASP8 and FADD-like apoptosis regulator	CFLAR	Hs.390736	pclust.2
222799_at	AK001606	HSPC049 protein	HSPC049	Hs.371722	pclust.2
211828_s_at	AF172268	TRAF2 and NCK interacting kinase	TNIK	Hs.34024	pclust.2
205968_at	NM_002252	potassium voltage-gated channel, delayed-rectifier, subfamily S, member 3	KCNS3	Hs.414489	pclust.2
226179_at	N63920	NA	NA	NA	pclust.2
212294_at	BG111761	guanine nucleotide binding protein (G protein), gamma 12	GNG12	Hs.431101	pclust.2
202386_s_at	NM_019081	limkain b1	LKAP	Hs.173524	pclust.2
225499_at	AW296194	KIAA1272 protein	KIAA1272	Hs.472285	pclust.2
205945_at	NM_000565	interleukin 6 receptor	IL6R	Hs.135087	pclust.2
221840_at	AA775177	protein tyrosine phosphatase, receptor type, E	PTPRE	Hs.127022	pclust.2
212076_at	AI701430	myeloid/lymphoid or mixed-lineage leukemia (trithorax homolog, Drosophila)	MLL	Hs.258855	pclust.2
217767_at	NM_000064	complement component 3	C3	Hs.529053	pclust.2
223269_at	BC004355	hypothetical protein MGC3200	MGC3200	Hs.9088	pclust.2

probeset	accession number	gene name	symbol	unigene	pclust
243729_at	AI457984	NA	NA	NA	pclust.2
231876_at	AL512757	tripartite motif-containing 56	TRIM56	Hs.521092	pclust.2
204014_at	NM_001394	dual specificity phosphatase 4	DUSP4	Hs.417962	pclust.2
217007_s_at	AK000667	a disintegrin and metalloproteinase domain 15 (metargidin)	ADAM15	Hs.312098	pclust.2
202552_s_at	NM_016441	cysteine-rich motor neuron 1	CRIM1	Hs.332847	pclust.2
225629_s_at	AI669498	zinc finger and BTB domain containing 4	ZBTB4	Hs.35096	pclust.2
218491_s_at	NM_014174	thymocyte protein thy28	THY28	Hs.13645	pclust.2
224215_s_at	AF196571	delta-like 1 (Drosophila)	DLL1	Hs.379912	pclust.2
235308_at	AW499525	zinc finger and BTB domain containing 20	ZBTB20	Hs.477166	pclust.2
200918_s_at	NM_003139	signal recognition particle receptor ('docking protein')	SRPR	Hs.368376	pclust.2
226981_at	AW002079	myeloid/lymphoid or mixed-lineage leukemia (trithorax homolog, Drosophila)	MLL	Hs.258855	pclust.2
212080_at	AV714029	myeloid/lymphoid or mixed-lineage leukemia (trithorax homolog, Drosophila)	MLL	Hs.258855	pclust.2
222453_at	AL136693	cytochrome b reductase 1	CYBRD1	Hs.221941	pclust.2
226160_at	AW138757	hexose-6-phosphate dehydrogenase (glucose 1-dehydrogenase)	H6PD	Hs.463511	pclust.2
208893_s_at	BC005047	dual specificity phosphatase 6	DUSP6	Hs.298654	pclust.2
225728_at	AI659533	NA	NA	NA	pclust.2
209270_at	L25541	laminin, beta 3	LAMB3	Hs.497636	pclust.2
211317_s_at	AF041461	CASP8 and FADD-like apoptosis regulator	CFLAR	Hs.390736	pclust.2
235651_at	AV741130	NA	NA	NA	pclust.2
202992_at	NM_000587	complement component 7	C7	Hs.78065	pclust.2
211495_x_at	AF114011	tumor necrosis factor (ligand) superfamily, member 13	TNFSF13	Hs.54673	pclust.2
210314_x_at	AF114013	tumor necrosis factor (ligand) superfamily, member 13	TNFSF13	Hs.54673	pclust.2
218084_x_at	NM_014164	FXYD domain containing ion transport regulator 5	FXYD5	Hs.333418	pclust.2
227026_at	AI016714	M-phase phosphoprotein, mpp8	HSMPP8	Hs.269654	pclust.2
222199_s_at	AK001289	bridging integrator 3	BIN3	Hs.546409	pclust.2
207836_s_at	NM_006867	RNA binding protein with multiple splicing	RBPMS	Hs.334587	pclust.2
225546_at	W68180	eukaryotic elongation factor-2 kinase	EEF2K	Hs.498892	pclust.2
224811_at	BF112093	NA	NA	NA	pclust.2
209468_at	AB017498	low density lipoprotein receptor-related protein 5	LRP5	Hs.6347	pclust.2

probeset	accession number	gene name	symbol	unigene	pclust
227983_at	AI810244	hypothetical protein MGC7036	MGC7036	Hs.488173	pclust.2
222862_s_at	BG169832	adenylate kinase 5	AK5	Hs.18268	pclust.2
214724_at	AF070621	DIX domain containing 1	DIXDC1	Hs.116796	pclust.2
1294_at	L13852	ubiquitin-activating enzyme E1-like	UBE1L	Hs.16695	pclust.2
203407_at	NM_002705	periplakin	PPL	Hs.192233	pclust.2
227507_at	BF593899	NA	NA	NA	pclust.2
215506_s_at	AK021882	ras homolog gene family, member I	ARHI	Hs.194695	pclust.2
226621_at	AI133452	fibrinogen, gamma polypeptide	FGG	Hs.546255	pclust.2
227040_at	AI655763	hypothetical protein LOC283506	LOC283506	Hs.507783	pclust.2
201814_at	AI300084	TBC1 domain family, member 5	TBC1D5	Hs.475629	pclust.2
222529_at	BG251467	mitochondrial solute carrier protein	MSCP	Hs.122514	pclust.2
231274_s_at	R92925	mitochondrial solute carrier protein	MSCP	Hs.122514	pclust.2
209499_x_at	BF448647	tumor necrosis factor (ligand) superfamily, member 12-member 13	TNFSF12-TNFSF13	Hs.54673	pclust.2
226728_at	BF056007	solute carrier family 27 (fatty acid transporter), member 1	SLC27A1	Hs.363138	pclust.2
213109_at	N25621	TRAF2 and NCK interacting kinase	TNIK	Hs.34024	pclust.2
212494_at	AB028998	tensin like C1 domain containing phosphatase	TENC1	Hs.6147	pclust.2
219563_at	NM_024633	chromosome 14 open reading frame 139	C14orf139	Hs.41502	pclust.2
209460_at	AF237813	4-aminobutyrate aminotransferase	ABAT	Hs.336768	pclust.2
211564_s_at	BC003096	PDZ and LIM domain 4	PDLIM4	Hs.424312	pclust.2
226597_at	AI348159	chromosome 19 open reading frame 32	C19orf32	Hs.76277	pclust.2
203941_at	NM_018250	hypothetical protein FLJ10871	FLJ10871	Hs.162397	pclust.2
208609_s_at	NM_019105	tenascin XB	TNXB	Hs.485104	pclust.2
201496_x_at	S67238	myosin, heavy polypeptide 11, smooth muscle	MYH11	Hs.460109	pclust.2
217732_s_at	AF092128	integral membrane protein 2B	ITM2B	Hs.446450	pclust.2
240120_at	H72914	NA	NA	NA	pclust.2
218380_at	NM_021730	NA	NA	NA	pclust.2
230492_s_at	BE328402	hypothetical protein KIAA1434	KIAA1434	Hs.472040	pclust.2
230472_at	AI870306	iroquois homeobox protein 1	IRX1	Hs.424156	pclust.2
224970_at	AA419275	nuclear factor I/A	NFIA	Hs.191911	pclust.2
204451_at	NM_003505	frizzled homolog 1 (Drosophila)	FZD1	Hs.94234	pclust.2
229817_at	AI452715	zinc finger protein 608	ZNF608	Hs.266616	pclust.2
229616_s_at	AU158463	hypothetical protein LOC196996	LOC196996	Hs.412093	pclust.2
218205_s_at	NM_017572	MAP kinase interacting serine/threonine kinase 2	MKNK2	Hs.515032	pclust.2
202304_at	NM_014923	fibronectin type III domain containing 3	FNDC3	Hs.508010	pclust.2
221795_at	AI346341	neurotrophic tyrosine kinase, receptor, type 2	NTRK2	Hs.494312	pclust.2

probeset	accession number	gene name	symbol	unigene	pclust
225776_at	AW205585	bromodomain adjacent to zinc finger domain, 2A	BAZ2A	Hs.314263	pclust.2
228496_s_at	AW243081	cysteine-rich motor neuron 1	CRIM1	Hs.332847	pclust.2
227438_at	AI760166	alpha-kinase 1	ALPK1	Hs.535761	pclust.2
223115_at	AK001674	cofactor required for Sp1 transcriptional activation, subunit 6, 77kDa	CRSP6	Hs.444931	pclust.2
221796_at	AA707199	neurotrophic tyrosine kinase, receptor, type 2	NTRK2	Hs.494312	pclust.2
225793_at	AW500180	hypothetical protein MGC46719	MGC46719	Hs.515748	pclust.2
201820_at	NM_000424	keratin 5 (epidermolysis bullosa simplex, Dowling-Meara/Kobner/Weber-Cockayne types)	KRT5	Hs.433845	pclust.2

Table C.1: Annotations for the 213 probesets selected after variable ranking (see Section 4.3.1).

Appendix D

Gene Ontology

The set of probesets used to construct the features (see Section 4.3.2) may contain tens or hundreds of genes. The common task is to translate this list of genes into a better understanding of the involved biological phenomena. Currently, this is done through a tedious combination of searches through the literature and a number of public databases. Fortunately useful tools (e.g. [Draghici et al., 2003]) allow to annotate automatically a list of genes.

To obtain some biological information, all genes were annotated according to known function using the Gene Ontology Consortium categories [Ashburner et al., 2000] : biological process, cellular component and molecular function. The GO consortium is setting a "dynamic controlled vocabulary that can be applied to all organisms even as knowledge of gene and protein roles in cells is accumulating and changing".

Such tools are used to obtain biological information about the probesets (as in [Lacroix et al., 2004]) and specifically in this thesis, the probesets used in the classifier.

Bibliography

Affymetrix (2002). *GeneChip Expression Analysis.*

Akritas, M. G. (1986). Bootstrapping the kaplan-meier estimator. *Journal of the American Statistical Association*, 81:1032–1038.

Akritas, M. G. (1994). Nearest neighbor estimation of a bivariate distribution under random censoring. *Annals of Statistics*, 22:1299–1327.

Allison, P. D. (1995). *Survival Analysis Using SAS: A Practical Guide.* SAS Institute Inc.

Amaldi, E. and Kann, V. (1998). On the approximation of minimizing non zero variables or unstaisfied relations in linear systems. *Theoretical Computer Science*, 209:237–260.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwoght, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene ontology: tool for the unfication of biology. the gene ontology consortium. *Nat Genet*, 25:25–29.

Bair, E. and Tibshirani, R. (2004). Semi-supervised methods to predict patient survival from gene expression data. *PLOS Biology*, 2(4):511–522.

Blum, A. and Langley, P. (1997). Selection of relevant features and examples in machine learning. *Artificial Intelligence*, 97(1-2):245–271.

Bolstad, B. M., Irizarry, R. A., Astrand, M., and TP, T. S. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193.

Brown, M., Grundy, W. N., Lin, D., Cristianini, N., Sugnet, C., Ares, M., and Haussler, D. (1999). Support vector machine classification of microarray gene expression data. University of California, Santa Cruz and University of Bristol.

Chang, J. C., Wooten, E. C., Tsimelzon, A., Hilsenbeck, S. G., Gutierrez, M. C., Elledge, R., Mohsin, S., Osborne, C. K., Chamness, G. C., Allred, D. C., and Connell, P. O. (2003). Gene expression profiling predicts therapeutic response to docetaxel in breast cancer patients. *Lancet*, 362:362–369.

Collett, D. (2003). *Modelling Survival Data in Medical Research.* Chapman and Hall, second edition edition.

Coombes, R. C. and Hall, E. (2004). A randomized trial of exemestane after two to three years of tamoxifen therapy in postmenopausal women with primary breast cance. *New England Journal Medecine*, 350(11):1081–1092.

Cox and Oakes, D. (1984). *Analysis of Survival Data*. Chapman and Hall (London).

Cox, D. R. (1972). Regression models and life tables. *Journal of the Royal Statistical Society Series B*, 34:187–220.

De Smet, F., Mathys, J., Marchal, K., TRhijs, G., De Moor, B., and Moreau, Y. (2002). Adaptive quality-based clustering of gene expression profiles. *Bioinformatics*, 18(5):735–746.

Draghici, S., Khatri, P., Bhavsar, P., Shah, A., Krawetz, S., and Tainsky, M. (2003). Onto-tools, the toolkit of the modern biologist: Pnto-express, onto-compare, onto-design and onto-translate. *Nucleic Acids Research*, 31(13):3775–3781.

Droesbeke, J. J. (1988). *Elements de Statistique*. Ellipses.

Duda, R. O., Hart, P. R., and Stork, D. G. (2001). *Pattern classification*. John Wiley and sons.

Dudoit, S., Fridlyand, J., and Speed, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, 97(457):77–87.

EBCT Collaborative Group (1998). Polychemotherapy for early breast cancer: an overview of the randomized trials. *Lancet*, 352:930–942. Early Breast Cancer Trialists' Collaborative Group.

Efron, B. (1977). The efficiency of cox's likelihood function for censored data. *Journal of the American Statistical Association*, 76:312–319.

Eifel, P., Axelson, J. A., Costa, J., Crowley, J., Curran, W. J., Deshler, A., Fulton, S., Hendricks, C. B., Kemeny, M., Kornblith, A. B., Louis, T. A., Markman, M., Mayer, R., and Roter, D. (2001). National institutes of health consensus development conference statement: Adjuvant therapy for breast cancer. *J. Natl Cancer Inst.*, 93(13):979–989.

Eisen, M., Spellman, P., Brown, P., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *PNAS*, 95:14863–14868.

Farewell, V. T. and Prentice, R. L. (1980). The approximation of partial likelihood with emphasis on case-control studies. *Biometrika*, 67:273–278.

Fdor, S. P., Read, J. L., Pirrung, M. C., Stryer, L., Lu, A. T., and Solas, D. (1991). Light-directed, spatially addressable parallel chemical synthesis. *Science*, 251(767-773).

Garfield, E. (1990). 100 most cited papers of all time. *Current Contents*.

Gautier, A., Moller, M., Frijs-Hansen, L., and Knudsen, S. (2005). Alternative mapping of probes to genes for affymetrix chips. *BMC Bioinformatics*, 5(111).

Gautier, L., Irizarry, R., Cope, L., and Boldstad, B. (2004). Description of affy. Technical report, Bioconductor.

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., R., A. J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J. Y., and Zhang, J. (2004). Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology*, 5:R80.

Goldhirsh, A., andR. D. Gelber, W. C. W., andB. Thurlimann, A. S. C., and Senn, H. J. (2003). Meeting highlights: Updated international expert consensus on the primary therapy of early breast cancer. *J. Clin. Oncol.*, 21(17):3357–3365.

Goldhirsh, A., Glick, J. H., Gelber, R. D., and Senn, H. (1998). Meeting highlights: International consensus panel on the treatment of primary breast cancer. *Journal of National Cancer Institute*, 90(1601-1608).

Goss, P. E. and Ingle, N. (2003). A randomized trial of letrozole in postmenopausal women after five years of tamoxifen therapy for early-stage breast cancer. *New England Journal of Medecine*, 349(19):1793–1802.

Greenwood, M. (1926). The errors of sampling of the survivorship tables. *Reports on Public Health and Statistical Subjects*, 33:1–26.

Gross, A. J. and Clark, V. A. (1975). *Survival Distributions: Reliability Applications in the Biomedical Sciences*. Wiley.

Gui, J. and Li, H. (2004). Penalized cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. *Center for Bioinformatics and Molecular Biostatistic, paper L1Cox.* http://repositories.cdlib.org/cbmb/L1Cox.

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182.

Haibe-Kains, B. (2004). Breast cancer diagnosis using microarray. Master's thesis, ULB.

Hartemink, A. J., Gifford, D. K., Jaakola, T. S., and Young, R. A. (2001). Maximum likelihood estimation of optimal scaling factors for expression array normalization. *SPIE BiOS*.

Hartigan, J. A. (1975). *Clustering Algorithms*. Wiley.

Hartmann, O., Samans, B., and Schafer, H. (2003). Low level analysis for affymetrix genechips: Normalization and quality control. Technical report, Insitiute of Medical Biometry and Epidemiology. Faculty of Medicine and Hospital, Philippe-University.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The elements of statistical learning.* Springer.

Heagerty, P. J., Lumley, T., and Pepe, M. S. (2000). Time-dependent roc curves for censored survival data and a diagnostic marker. *Biometrics*, 56:337–344.

Holder, D., Rauberras, R. F., Pikounis, V. B., Svetnik, V., and Shoper, K. (2001). Statistical analysis of high density oligonucleotide arrays: a safer approach. *Proceedings of the ASA Annual Meeting Atlanta, GA.*

Howell, A. and Cuzick, J. (2005). Results of the atac (arimidex, tamoxifen, alone or in combination) trial after completion of 5 years' adjuvant treatment for breast cancer. *Lancet*, 365(9453):60–62.

Hsieh, F. Y. (1995). A cautionary note on the analysis of extreme data with cox regression. *The American Statistician*, 49:226–228.

Irizarry, R. A., Bolstad, B. M., Collin, F., Cope, L. M., Hobbs, B., and Speed, T. P. (2003a). Summaries of affymetrix genechip probe level data. *Nucleic Acids Research*, 31(4):e15.

Irizarry, R. A., Bridget, H., Francois, C., Yasmin, D., Beazer-Barclay, Kristen, J., Antonellis, Uwe, S., and Speed, T. P. (2003b). Exploration, normalization, and summarization of hight density oligonucletotide array probe level data. *Bioinformatics.* in press.

Jansen, M., Foekens, J. A., van Staveren, I. L., Dirkzwager-Kiel, M. M., Ritstier, K., Look, M. P., van Gelder, M. E. M., Sieuwerts, A. M., Portengen, H., Dorssers, L. C., Jlijn, J., and Berns, M. (2005). Molecular clasification of tamoxifen-resistant breast carcinomas by gene expression profiling. *Journal of Clinical Oncology*, 23(4):732–740.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of American Statistical Asscoiation*, 53:457–451.

Kohavi, R. and John, G. (1997). Wrappers for feature subset selection. *AIJ*, 97(1-2):273–324.

Lacroix, M., Haibe-Kains, B., Laes, J. F., Hennuy, B., Lallemand, F., Gonze, I., Cardoso, F., Piccart, M., Leclercq, G., and Sotiriou, C. (2004). Gene regulation by phorbol 12-myristate 13-acetate (PMA) in two jighly different breast cancer cell lines. *Oncology Report*, 12(4):701–707.

Lacroix, M. and Leclercq, G. (2004). Relevance of breast cancer cell lines as models for breast tumors: an update. *Breast Cancer Res and Treat*, 415:530–536.

Lockhart, D. J., Dong, H., Byrne, M. C., Follette, M. T., Gallow, M. V., Chee, M. S., Mittmann, M., Wang, C., Kbayashi, M., Horton, H., and Brown, E. L. (1996). Expression monitoring by hybridization to high-density oligonucletoide arrays. *Nature Biotech.*, 14:1675–1680.

Ma, X. J., Wang, Z., Ryan, P. D., Isakoff, S. J., Barmettler, A., Fuller, A., Muir, B., Mohapatra, G., Salunga, R., Tuggle, J. T., Tran, Y., Tran, D., Tassin, A., Amon, P., Wang, W., Wang, W., Enright, E., Stecker, K., Estepa-Sabal, E., Smith, B., Younger, J., Balis, U., Michaelson, J., Bhan, A., Habion, K., Baer, T. M., Brugge, J., Haber, D. A., Erlander, M. G., and Sgroi, D. S. (2004). A two-gene expression ratio predicts clinical outcome in breast cancer patients treated with tamoxifen. *Cancer Cell*, 5:607–616.

Mason, S. J. and Graham, N. E. (1982). Areas beneath the relative operating characteristics (roc) and relative operating levels (rol) curves: Statistical significance and interpretation. *Q. J. R. Meteorol. Soc.*, 30:291–303.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Chapman and Hall.

Meier, P. (1975). Estimation of a distribution function from incomplete observations. *Perspectives in Probability and Statistics*, pages 67–87.

Michiels, S., Koscielny, S., and Hill, C. (2005). Prediction of cancer outcome with microarrays: a multiple radom validation strategy. *Lancet*, 365:488–492.

Mitchell, T. (1997). *Machine Learning*. McGraw.

Naef, F., Lim, D. A., and Magnasco, M. O. (2001). rom features to expression: High desnity oligonucleotide array analysis revisited. Technical report, Institut fur Hydromechanik und Wasserwirtschaft.

Paik, S., Shak, S., Tang, G., Kim, C., Bakker, J., Cronin, M., Baehner, F. L., Walker, M. G., Watson, D., Park, T., Hiller, W., Fisher, E. R., Wickerham, D. L., Bryant, J., and Wolmark, N. (2004). A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *New England Journal of Medecine*, (351):2817–2826.

Ploner, A., Miller, L. D., Hall, P., Bergh, J., and Pawitan, Y. (2005). Correlation test to assess low-level processing of high-desnity oligonucletide microarray data. *BMC Bioinformatics*.

R Development Core Team (2005). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rossi, P. H., Berk, R. A., and Lenihan, K. J. (1980). *Money, Work and Crime: Some Experimental Results*. New Yord Academic Press Inc.

Sheden, K., Chen, W., Kuick, R., Ghosh, D., Macdonald, J., Cho, K. R., Giordano, T. J., Gruber, S. B., Fearon, E. R., Taylor, J. M., and Hanash, S. (2005). Comparison of seven methods for producing affymetrix expression score based on false discovery rate in disease profiling data. *BMC Bioinformatics*, 6(26).

Shipp, M. A., Ross, K. N., and Tamayo, P. (2002). Diffuse large b-cell lymphoma outcome prediction by gene-expression profiling. *Nature Medicine*, 8:68–74.

Sotiriou, C., Wirapati, P., Loi, S., Harris, A., Bergh, J., Smeds, J., Farmer, P., Praz, V., Haibe-Kains, B., Lallemand, F., Buyse, M., Piccart, M., and Delorenzi, M. (2005). Gene expression profiling in breast cancer challenges the existence of intermediate histological grade. *submitted*.

Therneau, T. M. and Grambsch, P. M. (2000). *Modeling Survival Data: Extending the Cox Model*. Springer.

Tibshirani, R. (1997). The lasso method for variable selection in the cox model. *Statistics in Medecine*, 16:385–395.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison.

van de Vijver, M. J., He, Y. D., van't Veer, L., Dai, H., Hart, A. M., Voskuil, D. W., Schreiber, G. J., Peterse, J. L., Roberts, C., Marton, M. J., Parrish, M., Atsma, D., Witteveen, A., Glas, A., Delahaye, L., van der Velde, T., Bartelink, H., Rodenhuis, S., Rutgers, E. T., Friend, S. H., and Bernards, R. (2002). A gene expression signature as a predictor of survival in breast cancer. *The new England Journal of Medecine*, 347(25):1999–2009.

van't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A., Mao, M., Peterse, H. L., van der Kooy, K., Marton, M. J., Witteveen, A. T., Schreiber, G. J., Kerkhiven, R. M., Roberts, C., Linsley, P. S., Bernards, R., and Friend, S. H. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415:530–536.

Vittinghof, E., Glidden, D. V., Shiboski, S. C., and McCulloch, C. E. (2005). *Regression Methods in Biostatistics: Linear, Logistic, Survival and Repeated Measures Models.* Springer.

Yu, H., Luscombe, N. M., Gian, J., and Gerstein, M. (2003). Genomic analysis of gene expression relationships in transcriptional regulatory networks. *Trend in Genetics*, 19(8):422–427.