UNIVERSITE LIBRE DE BRUXELLES
FACULTE DES SCIENCES
DEPARTEMENT D'INFORMATIQUE

Identification and Assessment of Gene Signatures in Human Breast Cancer

Thèse présentée par

Benjamin Haibe-Kains

En vue de l'obtention du grade de

Docteur en Sciences

Avril 2009

This thesis has been written under the supervision of Prof. Gianluca Bontempi and Prof. Christos Sotiriou.

The members of the Jury are:

- Prof. Gianluca Bontempi (Université Libre de Bruxelles, Belgium)
- Prof. Christos Sotiriou (Institut Jules Bordet, Université Libre de Bruxelles, Belgium)
- Prof. Tom Lenaerts (Université Libre de Bruxelles, Belgium)
- Prof. Christine Decaesteker (Université Libre de Bruxelles, Belgium)
- Prof. Jacques van Helden (Université Libre de Bruxelles, Belgium)
- Prof. Hughes Bersini (Université Libre de Bruxelles, Belgium)
- Prof. Yves Moreau (Katholieke Universiteit Leuven, Belgium)
- Prof. Mauro Delorenzi (Swiss Institute of Bioinformatics, Switzerland)

To my future child.

Résumé

Cette thèse concerne le développement de techniques d'apprentissage (*machine learning*) afin de mettre au point de nouveaux outils cliniques basés sur des données moleculaires. Nous avons focalisé notre recherche sur le cancer du sein, un des cancers les plus fréquemment diagnostiqués. Ces outils sont développés dans le but d'aider les médecins dans leur évaluation du devenir clinique des patients cancéreux (cf. le pronostique).

Les approches traditionnelles d'évaluation du pronostique d'un patient cancéreux se base sur des critères clinico-pathologiques connus pour être prédictifs de la survie. Cette évaluation permet aux médecins de décider si un traitement est nécessaire après l'extraction de la tumeur. Bien que les outils d'évaluation traditionnels sont d'une aide importante, les cliniciens sont conscients de la nécessité d'améliorer de tels outils.

Dans les années 90, de nouvelles technologies à haut-débit, telles que le profilage de l'expression génique par biopuces à ADN (*microarrays*), ont été mises au point afin de permettre aux scientifiques d'analyser l'expression de l'entièreté du génôme de cellules cancéreuses. Ce nouveau type de données moléculaires porte l'espoir d'améliorer les outils pronostiques traditionnels et d'approfondir nos connaissances concernant la génèse du cancer du sein. Cependant ces données sont extrêmement difficiles à analyser à cause (i) de leur haute dimensionalité (plusieurs dizaines de milliers de gènes pour seulement quelques centaines d'expériences); (ii) du bruit important dans les mesures; (iii) de la collinéarité entre les mesures dûe à la co-expression des gènes.

Depuis 2002, des études comparatives à grande échelle ont permis d'identifier les méthodes performantes pour l'analyse de groupements et la classification de données microarray, négligeant l'analyse de survie pertinente pour le pronostique dans le cancer du sein. Pour pallier ce manque, cette thèse présente une méthodologie originale adaptée à l'analyse de données microarray et de survie afin de construire des modèles pronostiques performants et robustes.

En termes d'applications, nous montrons que cette méthodologie, utilisée en combinaison avec des connaissances biologiques *a priori* et de nombreux ensembles de données publiques, a permis d'importantes découvertes. En particulier, il résulte de la recherche presentée dans cette thèse, le développement d'un modèle robuste d'identification des sous-types moléculaires du cancer du sein et de plusieurs signatures géniques améliorant significativement l'état de l'art au niveau pronostique.

Summary

This thesis addresses the use of machine learning techniques to develop clinical predictive tools for breast cancer using molecular data. These tools are designed to assist physicians in their evaluation of the clinical outcome of breast cancer (referred to as prognosis).

The traditional approach to evaluating breast cancer prognosis is based on the assessment of clinico-pathologic factors known to be associated with breast cancer survival. These factors are used to make recommendations about whether further treatment is required after the removal of a tumor by surgery. Treatment such as chemotherapy depends on the estimation of patients' risk of relapse. Although current approaches do provide good prognostic assessment of breast cancer survival, clinicians are aware that there is still room for improvement in the accuracy of their prognostic estimations.

In the late 1990s, new high throughput technologies such as the gene expression profiling through microarray technology emerged. Microarrays allowed scientists to analyze for the first time the expression of the whole human genome ("transcriptome"). It was hoped that the analysis of genome-wide molecular data would bring new insights into the critical, underlying biological mechanisms involved in breast cancer progression, as well as significantly improve prognostic prediction. However, the analysis of microarray data is a difficult task due to their intrinsic characteristics: (i) thousands of gene expressions are measured for only few samples; (ii) the measurements are usually "noisy"; and (iii) they are highly correlated due to gene co-expressions. Since traditional statistical methods were not adapted to these settings, machine learning methods were picked up as good candidates to overcome these difficulties. However, applying machine learning methods for microarray analysis involves numerous steps, and the results are prone to overfitting. Several authors have highlighted the major pitfalls of this process in the early publications, shedding new light on the promising but overoptimistic results.

Since 2002, large comparative studies have been conducted in order to identify the key characteristics of successful methods for class discovery and classification. Yet methods able to identify robust molecular signatures that can predict breast cancer prognosis have been lacking. To fill this important gap, this thesis presents an original methodology dealing specifically with the analysis of microarray and survival data in order to build prognostic models and provide an honest estimation of their performance. The approach used for signature extraction consists of a set of original methods for feature transformation, feature selection and prediction model building. A novel statistical framework is presented for performance assessment and comparison of risk prediction models.

In terms of applications, we show that these methods, used in combination with *a priori* biological knowledge of breast cancer and numerous public microarray datasets, have resulted

in some important discoveries. In particular, the research presented here develops (i) a robust model for the identification of breast molecular subtypes and (ii) a new prognostic model that takes into account the molecular heterogeneity of breast cancers observed previously, in order to improve traditional clinical guidelines and state-of-the-art gene signatures.

Acknowledgments

Une thèse ne se résume pas qu'à l'aboutissement d'une recherche scientifique, d'expériences et de rédaction. C'est une tranche de vie, incluant tout son entourage. Je tiens à remercier toutes ces personnes, qui à travers leur amour, leur amitié, leur soutien et leurs idées, ont contribué à l'aboutissement du présent travail.

Tout d'abord, ma famille, tout particulièrement

> *Ma mère à qui je dois tout (et plus encore),*
>
> *Mes pères qui ont contribués chacun à leur manière à faire de moi ce que je suis devenu notamment au niveau professionnel,*
>
> *Mes nombreux frères et soeurs qui m'ont énormément appris (bien plus qu'ils ne le sauront jamais),*
>
> *Mon grand-père qui m'a appris l'importance du passé pour apprendre (il m'a sans doute inculqué mes premières notions d'apprentissage supervisé),*
>
> *Ma grand-mère, qui m'a donné le goût du voyage (serait-elle à l'origine de mon projet de PostDoc aux USA?).*

Ensuite, mes amis pour leur soutien inconditionel à toutes les étapes de ma vie. Parmi eux, j'ai rencontré le bonheur à plus d'un titre. Une jeune femme notamment, si pleine de vie et d'amour qu'elle a accepté de lier nos vies à jamais.

After my family and friends, I would like to thank those who contributed directly to this thesis:

> Gianluca Bontempi, who has followed me since I received my degree in computer science and who initiated the collaboration with Christos Sotiriou. His help has been invaluable for my research.
>
> Christos Sotiriou, who has allowed me to do my training in his lab and to stay there. Like Gianluca, his passion for research is a great source of motivation. I thank him also for his confidence, which allows me to continue my research in spite of the tremendous volume of work in the lab.
>
> Christine Desmedt, Sherene Loi, Asa Wirapati, Mauro Delorenzi and Marc Buyse, with whom the collaboration has been more than a pleasure.

All my colleagues of the Machine Learning Group (Yann-Aël Le Borgne, Patrick Meyer, Olivier Caelen, Kevin Kontos, Abhilash Miranda, Catharina Olsen, Olivier Cailloux, Mathieu Van der Haegen, Mehdi Moussaid, and Benjamin Tshibasu-Kabeya) and the Functional Genomics Unit (Virginie Durbecq, Françoise Lallemand, Ghizlane Rouas, Françoise Rothé, Carole Equeter, Benjamin Bopp, Samira Majjaj, Jérôme Toussaint, Mounia Bouzeghrane, Carole Chaboteaux, Michail Ignatiadis, Naïma Kheddoumi, Sandeep Singhal, Carine Vanderstraeten), for their enthusiasm and the great discussions we have had.

All of my teachers from the Computer Science department and the Bioinformatics master, who have opened my mind to such interesting fields of research.

Carolyn Straehle for her editorial comments and her numerous encouragements throughout the writing process.

I have been lucky to benefit, beside my colleagues, jury committee and my supervisors, from the careful proofreading of Raymond Devillers. I thank him for his support since my degree in Computer Science and for his interest in my research.

The members of my jury who have accepted to comment this work, namely Prof. Tom Lenaerts (ULB), Christine Decaesteker (ULB), Jacques van Helden (ULB), Hughes Bersini (ULB), Yves Moreau (Katholieke Universiteit Leuven) and Mauro Delorenzi (Swiss Institute of Bioinformatics, Switzerland).

Again, my wife, Olivia, for putting up with me, even when I live only for my work.

Financial Support

The work presented in this thesis was supported by the Belgian National Funds for Scientific Research (FNRS) through a Télévie grant.

Declaration

The thesis has been composed by the author himself and contains original work of his own execution. Some of the reported work has been done in collaboration with the Bioinformatics Core Facility, headed by Prof. Mauro Delorenzi, at the Swiss Institute of Bioinformatics, Lausanne, Switzerland.

In particular, I would like to acknowledge Dr. Pratyaksha Wirapati and Prof. Mauro Delorenzi for the fruitful collaborative work presented in Sections 4.1.3.1 and 4.2.2 whose related experimental findings are reported in Sections 5.2.1, 5.2.3, and 5.4.1.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This interdisciplinary work concerns the development of original predictive tools in medicine based on molecular data. In particular, we focus our research on machine learning methods for breast cancer prognostication from microarray data. Our approach stresses the robustness of the predictive models, their biological interpretation, and their application to different microarray technologies.

In recent decades, we have witnessed an increased incidence of cancer, rendering cancer one of today's major public health issues. Currently, *breast cancer* is the most frequently diagnosed malignancy in women in the Western world.

From the 1990s, new high throughput technologies have emerged and enabled the study of disease at the molecular level. These technologies, such as gene expression profiling through *microarrays*, carry with them the hope of bringing new insights to cancer biology and improving current tools for cancer management.

In breast cancer, two main issues are *prognostication* and *prediction* of therapy benefit. An accurate prognostic tool would enable doctors to anticipate the prospect of remission from the usual course of disease, and therefore to spare patients from unnecessary anti-cancer treatments (and their concomitant adverse side effects). An accurate predictive tool would enable doctors to anticipate the response or resistance of a patient to an anti-cancer treatment, and therefore to select the most suitable treatment available.

Traditional clinical tools for prognostication and prediction are based on a small set of variables routinely measured in the clinic. These tools are far from perfect and much progress is needed to yield accurate risk predictions. Clinical investigators have rapidly harnessed the great potential of the high throughput technologies, not only for gaining new insights into cancer biology, but also to improve these traditional clinical tools.

The objective of this thesis is to develop original prognostic and predictive tools using molecular data generated by high throughput technologies in order to improve traditional clinical tools. The complexity of the data and the interdisciplinary context of the problem make this task extremely challenging.

The contributions of the thesis will be described in terms of medical implications, biological findings and methodology. While the novel methods we developed will be covered in detail, we will present the medical implications and the biological knowledge we used or generated to a somewhat more limited extent.

The Introduction describes the biomedical and bioinformatics contexts of the thesis. The

medical questions of interest, the technology used to generate the data, and the state of bioinformatics research at the time work on this thesis was begun are introduced in the following sections. This chapter ends with a brief description of the contributions of this thesis and the various notations used in it.

1.1 Breast Cancer

Breast cancer is a global public health issue. It is the most frequently diagnosed malignancy in women in the Western world and the commonest cause of cancer death in European and American women. According to estimates in 2002, there were 1,151,298 new cases of breast cancer diagnosed, 410,712 deaths caused by breast cancer, and more than 44 million women living with breast cancer worldwide [Veronesi et al., 2005]. In Europe, one out of eight to ten women, depending on the country, will develop breast cancer during her lifetime [Parkin et al., 2001].

Thanks to the routine use of screening mammograms in developed countries, more and more women diagnosed with breast cancer are detected at an early stage (early breast cancer, small tumors and absence of lymph node invasion). Surgery is the primary treatment in the majority of cases, alone or in combination with radiotherapy. Despite early detection, up to 50% of these women will develop *distant metastasis*, i.e. development of new tumors in different organs. Metastatic breast cancer is unfortunately incurable. As a result, since the mid 1980s, randomized trials of adjuvant systemic therapy have been conducted in an effort to reduce the rate of recurrence and to prolong the survival of patients with operable disease [EBCTG, 2005].

Due to the importance of breast cancer for public health, this field has been the subject of intense research for decades. Moreover, new high throughput technologies, such as gene expression profiling, became readily available at the end of the 1990s, providing powerful tools to study and fight this disease.

1.1.1 Biological Insights Through Gene Expression Profiling

Gene expression profiling, through microarray-based technology, is a powerful tool with which to draw up a genetic portrait of a biological sample (e.g. a tumor sample). Contrary to traditional molecular and genetic profiling methods that focus on a few genes at a time, microarray techniques allow for the simultaneous evaluation of the expression of thousands of genes. Clinical investigators rapidly harnessed the great potential of this technology for gaining new insights into cancer biology.

Since the research carried out in this thesis involves gene expression data, this technology is described in greater detail in Section 2.1. We present below the early studies in breast cancer biology using this technology at the time research for this thesis was begun.

Clinicians have long recognized that breast tumors exhibit different natural histories and responses to various treatments. Nevertheless, traditional *histo-pathological* characteristics, i.e. microscopic examination of the diseased tissues anatomy, are unable to capture the biologic heterogeneity of these tumors. Many early studies attempted to identify subtypes of breast tumors using gene expression data without taking into account *a priori* biological knowledge [Perou et al., 2000; Sorlie et al., 2001, 2003; Sotiriou et al., 2003]. These studies

used a clustering method (unsupervised approach, see Section 2.2) to consistently show that (i) genes related to estrogen receptor (ER) and human epidermal growth factor receptor 2 (HER2) signaling pathways[1] have the strongest association with the gene expression profile of breast tumors; (ii) these tumors can be grouped into at least four subtypes, the basal-like (ER-/HER2-), the HER2+ and two luminal (ER+/HER2-) subtypes characterized mainly by different expression levels of proliferation genes; and (iii) each subtype exhibits distinct clinical outcomes. Interestingly, these studies also revealed that clinically relevant indicators such as menopausal status, tumor size and nodal status were not associated with distinct gene expression profiles. This class discovery has been useful to highlight that breast cancers are a heterogeneous group of diseases, and it has helped to better understand breast tumor biology. However, it is difficult at this stage to use these results to improve breast cancer prognostication or prediction (Sections 1.2 and 1.3), and this is due to the difficulty related to fitting such a clustering model and using it for new cases [Pusztai et al., 2006]. Moreover, although these methods have been effective at highlighting biological differences between tumors, they do not look for the differentially expressed genes with respect to the clinical outcome of interest. Thus, they are ill-adapted to identifying relevant prognostic and/or predictive genes.

1.2 Prognostication

The goal of prognostication is to predict the survival of a patient, or her risk to develop metastases without treatment (Figure 1.1). Specifically, prognosis attempts to predict the prospect of remission of a breast cancer patient from the usual course of disease after the initial surgery. This information is extremely important because it assists oncologists in determining which breast cancer patients require chemo-, hormono- or other systemic therapies, and which women can safely be treated with radiotherapy alone.

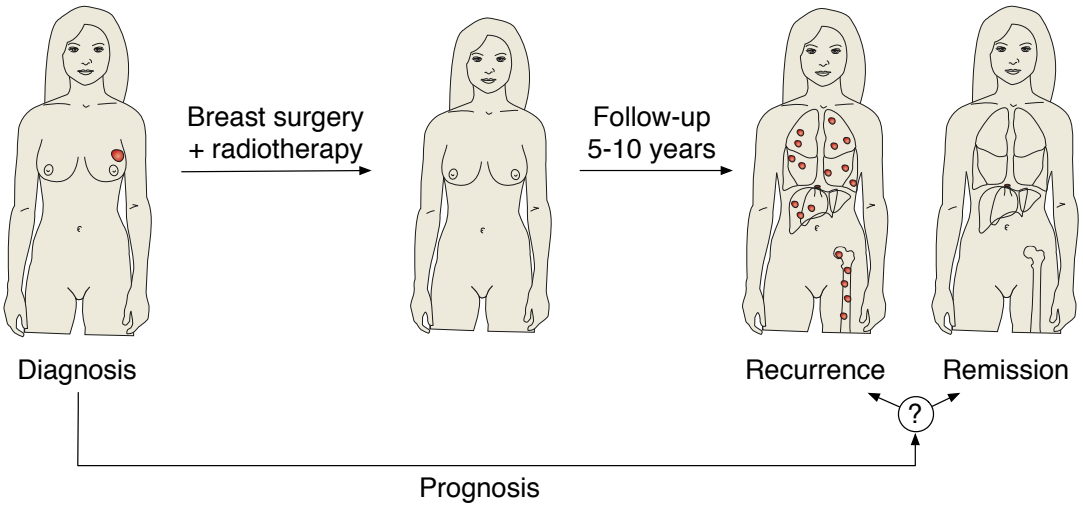


Figure 1.1: Breast cancer prognostication. Figure adapted from [Sotiriou and Piccart, 2007].

[1] See Section 3.1 for a description of the main biological processes involved in breast cancer.

1.2.1 Traditional Approach

There are several clinical variables commonly used for breast cancer prognosis, as depicted in Figure 1.2. The risk of recurrence is primarily determined by the age of the patient, nodal status, tumor size, histological grade, the expression status of the hormonal receptors, i.e. the estrogen (ER) and the progesterone receptors (PgR) as quantified by immunohistchemistry (IHC), and the expression (IHC) or the gene amplification (fluorescence In situ hybridization, FISH) status of the HER2 oncogene. These clinical variables can provide prognostic information and are summarized in clinical guidelines, such as the National Institute of Health (NIH; [Eifel et al., 2001]) in the USA or the St Gallen consensus criteria [Goldhirsh et al., 2003] in Europe in order to assist clinicians and patients in adjuvant therapy decision-making.

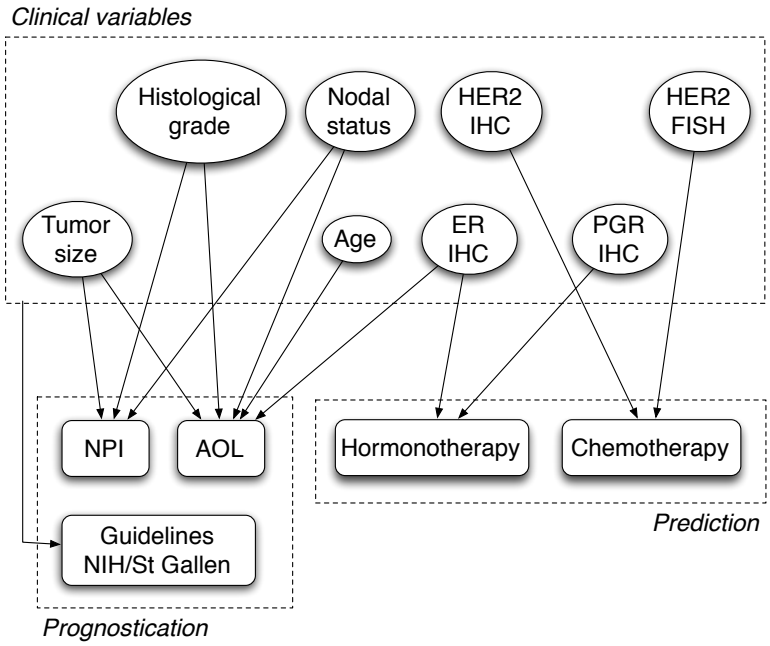


Figure 1.2: Traditional prognostic and predictive tools for breast cancer used in the clinic before the advent of new high throughput technologies such as gene expression profiling. The clinical guidelines for prognostication use all the clinical variables available.

We illustrate below the use of a clinical variable for breast cancer prognostication through histological grade. Histological grade [Scarff and Torloni, 1968] is a well-known histo-pathological parameter routinely used in the clinic to measure tumor differentiation, i.e. how much tumor cells look like the normal tissue from which they originated:

- Histological grade 1, or well-differentiated tumor cells, look very much like normal, nearby breast tissue.
- Histological grade 2, or moderately differentiated tumor cells, exhibit an intermediate differentiation stage between well and poorly differentiated tumor cells.
- Histological grade 3, or poorly differentiated tumor cells, show very few similarities to normal breast tissue.

Histological grade is known to be highly prognostic in breast cancer [Elston and Ellis, 1991]. Patients having a histological grade 1 tumor exhibit better survival than patients having a histological grade 3 tumor. However, clinicians face a tremendous problem with patients who have histological grade 2 tumors, because these tumors, which represent 30% to 60% of breast cancer cases, are the major source of inter-observer discrepancy, uncertainty in histological grade determination, and exhibit intermediate survival, making treatment decision-making for these patients a great challenge [Singletary et al., 2002; Perez et al., 2006]. So, the use of histological grade is not sufficient to predict precisely the clinical outcome of a breast cancer patient.

To reduce uncertainty in prognosis, these clinical variables can also be combined into multivariable outcome prediction models, like Adjuvant! Online (AOL; [Olivotto et al., 2005]) and the Nottinghman Prognostic Index (NPI; [Galea et al., 1992]). These tools use some of the clinical variables to estimate the risk of recurrence of breast cancer patients (Figure 1.2). However, risk estimation based on these guidelines or prognostic models is far from perfect and much progress is needed before it will be possible to clearly identify those patients, especially with early (node-negative, i.e. nodal status equal to 0) breast cancer, who would really need adjuvant systemic therapy [Isaacs et al., 2001; Sotiriou and Piccart, 2007]. As a result, many women are prescribed adjuvant chemotherapy who probably would have had excellent long term outcomes without it, exposing them to the potential adverse effects of chemotherapy such as cardiac dysfunction, second malignancies and premature menopause. Therefore, better prognostic tools could avoid the adverse side effects of adjuvant therapies, as well as the high costs of such treatments.

1.2.2 Gene Expression Profiling Approach

During the last two decades, several clinical and pathological parameters have been used to evaluate the prognosis of breast cancer patients. Although different tools have been developed to assist clinicians in selecting patients who should receive adjuvant therapy (Figure 1.2), such as the St Gallen consensus criteria [Goldhirsh et al., 2003], the NIH guidelines [Eifel et al., 2001], the Nottingham Prognostic Index [Galea et al., 1992] or Adjuvant! Online [Olivotto et al., 2005], it still remains a challenge to distinguish those patients who would really need adjuvant systemic therapy from those who could be spared such treatment.

Clinical investigators rapidly harnessed the great potential of gene expression profiling, not only for gaining new insights into cancer biology (Section 1.1.1), but also as a powerful prognostic tool. Unlike the traditional variables routinely measured in the clinic (Figure 1.2) which are limited to few, sometimes subjective, measurements, this technology enables the quantitative measurement of thousands of gene expressions in parallel (Section 2.1), making possible the development of prognostic models with numerous molecular markers.

In order to develop a more accurate tool for early breast cancer prognosis, the Netherlands Cancer Institute (NKI) conducted a comprehensive, genome-wide assessment of gene expression profiling [van't Veer et al., 2002]. By using Agilent microarray technology (see Section 2.1 for details), they identified the genes differentially expressed between two groups of patients that differ in their survival. The low-risk group included patients who had not developed distant metastases within the first five years after diagnosis, a result that contrasted with the high-risk group. The NKI group refined the set of relevant genes and built a risk prediction model with 70 prognostic genes (denoted by GENE70). This set of genes (also known as *gene signature*) included mainly genes involved in the cell cycle, invasion, metas-

tasis, angiogenesis and signal transduction. This gene signature was then validated on a larger set of patients, including both node-negative and node-positive breast tumors in treated and untreated patients from the same institution [van de Vijver et al., 2002], and consequently proved to be the strongest predictor for distant metastasis-free survival (DMFS, see Section 5.1 for details), independently of several clinical prognostic indicators described in Figure 1.2. To assess the clinical relevance of the GENE70 signature, the authors compared its performance to that of current commonly used breast cancer risk classification criteria, i.e. the National Institute of Health (NIH) consensus [Eifel et al., 2001], and the St Gallen guidelines [Goldhirsh et al., 2003]. Citing as evidence the excellent disease free outcomes for patients at five years, the NKI group found that the GENE70 signature, compared to the NIH and St Gallen classifications, was better at predicting which patients should have been spared adjuvant chemotherapy (low risk) and which patients should have been prescribed adjuvant chemotherapy (high-risk). The authors concluded that the GENE70 signature could outperform current clinical risk classifications and therefore could significantly impact on breast cancer management by sparing some women from over-treatment and the unnecessary toxicity of chemotherapy.

Using a similar approach a few years later, Erasmus Medical Center and Veridex identified a prognostic gene signature (denoted by GENE76) that could be used to predict the development of distant metastases within the first five years after diagnosis in early (node-negative) breast cancer patients who did not receive systemic treatment [Wang et al., 2005]. In contrast to van't Veer et al., this study used Affymetrix microarray technology (see Section 2.1 for details) to build a risk prediction model that considered ER-positive patients separately from ER-negative patients. This decision was based on the assumption that the mechanisms for disease progression could differ for these two ER-based subgroups of breast cancer patients. Similarly to the GENE70 signature, when compared to the classification results of two conventional sets of consensus guidelines, St. Gallen and NIH, the GENE76 signature better identified the low-risk patients not needing treatment.

By using gene expression profiling to develop gene signatures that are advantageous when compared to clinical guidelines, we could therefore significantly reduce the number of patients subject to unnecessary treatment. This would ultimately also translate into savings in cost and health resources, without sacrificing long term clinical outcome. However, a careful validation of the gene expression profiling technology and prognostic gene signatures is required before bringing this predictive tool into day-to-day clinical practice.

1.3 Prediction

The use of systemic adjuvant treatments has increased in the last ten years, with the objective of prolonging the survival of breast cancer patients. New treatments are continually being developed in order to target specifically the cancer cells and to reduce toxicity for the individual. The goal of prediction is to predict the response of a breast cancer patient to a treatment.

There exist two settings for breast cancer prediction: the *adjuvant* (Figure 1.3) and the *neoadjuvant* (Figure 1.4) settings [Mauri et al., 2005]. The adjuvant setting is similar to the prognostication illustrated in Figure 1.1, except that the patients are prescribed a therapy. In the neoadjuvant setting, the situation is more complex. First, a biopsy of the breast tumor

is taken at diagnosis, before the neoadjuvant therapy. Second, breast surgery is carried out to remove the tumor and to assess whether the tumor was affected by the treatment (e.g. decrease in tumor size). A pathological complete response (pCR) is then defined as the complete disappearance of tumor cells in the breast and the axillary lymph nodes and it has been shown that a pathological complete response is associated with excellent long-term survival. In this case, only the response or the resistance to the treatment is analyzed, leaving aside the issue of the survival of the patients. In this thesis, we will focus on the adjuvant setting to study prediction in breast cancer.

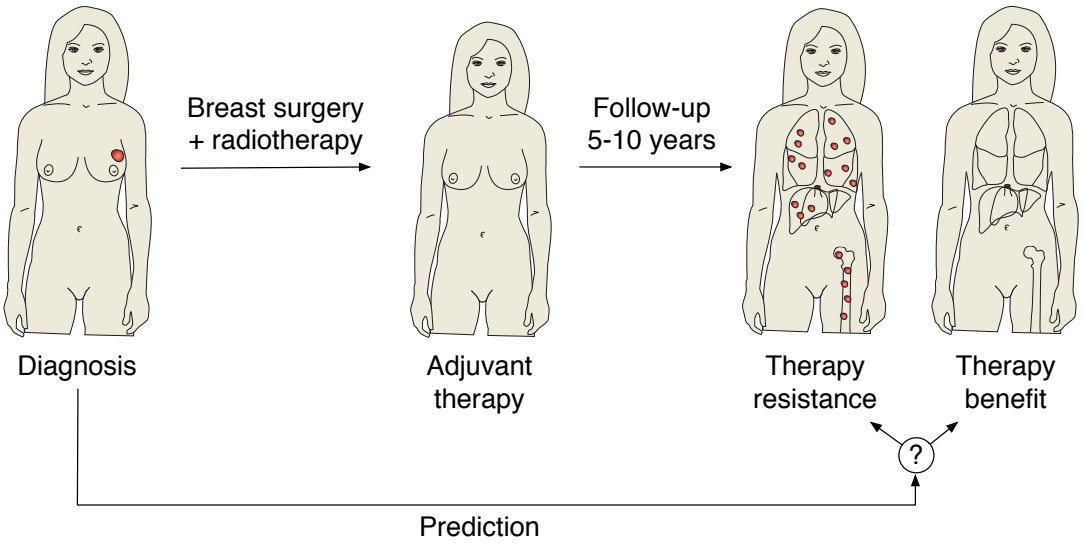


Figure 1.3: Adjuvant setting for breast cancer prediction. Figure adapted from [Sotiriou and Piccart, 2007].

Although accurate breast cancer prognosis allows for identification of the patients needing treatment, clinicians also need to know which therapy will benefit the individual patient most. Indeed, only a proportion of patients will respond to a particular treatment, whereas most will experience its adverse side effects. Moreover, the current over-treatment of patients results in major expenses for individuals and society.

1.3.1 Traditional Approach

Currently, there exist few tools for prediction. For instance, the expression status of the hormonal receptors (ER and PGR) and the expression/gene amplification status of the HER2 oncogene are used to define the subset of individuals who may benefit from hormono- and chemotherapy, respectively (Figure 1.2).

Despite the existence of the tools described above, current prediction models need to be improved, since the accuracy of these tools is poor [Sotiriou and Piccart, 2007; Lonning et al., 2007]. Numerous attempts have been made to identify prognostic groups based on other pathological characteristics, mainly lymphovascular invasion or proliferation markers such as S-phase fraction, which might better reflect tumor biology and serve as prognostic and/or predictive markers that may aid in treatment decision making in the adjuvant set-

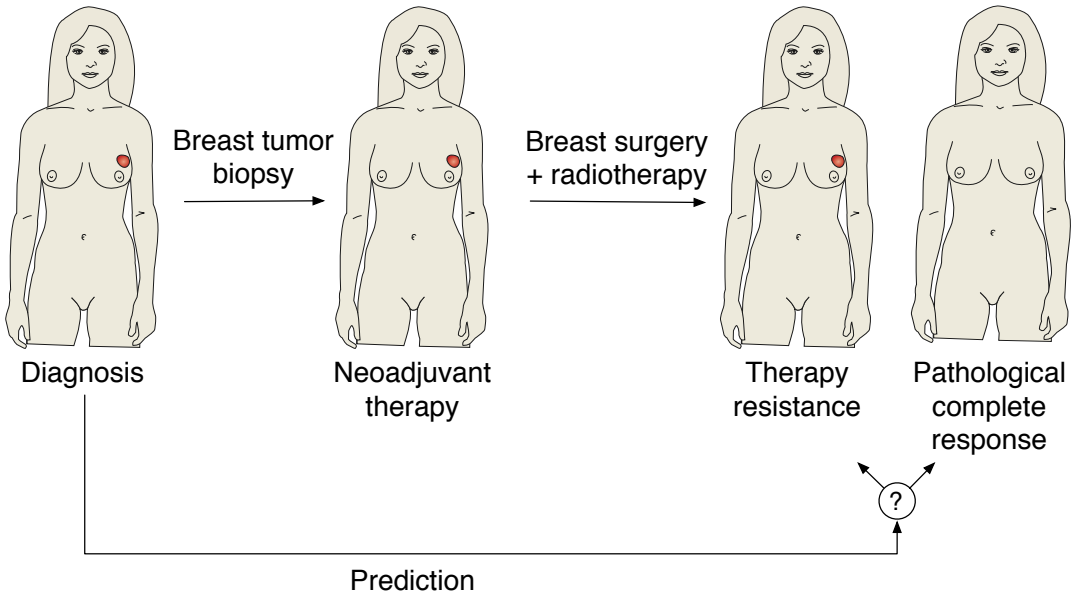


Figure 1.4: Neoadjuvant setting for breast cancer prediction. Figure adapted from [Sotiriou and Piccart, 2007].

ting (reviewed in [Colozza et al., 2005]). In addition, a variety of molecular tumor markers have been studied both in the laboratory and in the clinical settings for their ability to predict response to treatment [reviewed in [Colozza et al., 2005]]. Unfortunately, the studies examining the clinical utility of these tumor markers have usually used small, heterogeneous, retrospective patient series, often with insufficient power to draw robust conclusions; moreover, they have not been reported in a detailed enough fashion to provide information for the reproduction and external validation of results [McShane et al., 2005]. There is also a lack of well-designed, prospective clinical trials addressing the clinical utility of such markers.

Beyond this, given the complexity of breast cancer and the huge diversity in molecular pathways dissected by basic research scientists [Konecny et al., 2004], isolated markers might not be sufficient to predict response or resistance to treatment, and a comprehensive view of the disease is needed.

These limitations have driven breast cancer research to develop more accurate molecular predictors of clinical outcome and response to various anti-cancer therapies using a multi-marker approach with the help of the quantitative gene expression profiling technologies.

1.3.2 Gene Expression Profiling Approach

Similarly to the use of gene expression profiling for prognostication, this technology carries the hope to improve current breast cancer predictive tools by the development of new predictive models using numerous molecular markers.

The idea of applying gene expression profiling to identify new predictive signatures has only been applied in a limited number of studies that have included patients treated with a number of different standard systemic therapies.

Benefit of hormonotherapy Two predictors have been identified for the benefit of tamoxifen therapy, the most widely used type of hormonotherapy.

Ma et al. developed a gene signature predictive of relapse free survival (RFS, see Section 5.1 for details) from 60 patients treated with adjuvant tamoxifen [Ma et al., 2004]. The signature was reduced to a two-gene expression ratio, HOXB13 versus IL17RB, transposed onto a technology based on polymerase chain reaction (PCR) and validated on an independent series using standard, formalin-fixed paraffin embedded tissue.

Similarly, Paik et al. developed a 16-gene assay on formalin-fixed paraffin-embedded samples called the recurrence score (denoted by ONCOTYPE), which can predict the risk of recurrence in patients receiving adjuvant tamoxifen [Paik et al., 2004]. The ONCOTYPE signature can be used to estimate the probability of recurrence at 10 years or can be used to classify patients into low-, intermediate-, or high-risk categories. ONCOTYPE performance for distant metastasis prediction was assessed retrospectively in 668 patients with ER-positive, node-negative breast cancers treated with tamoxifen who were enrolled in the National Surgical Adjuvant Breast and Bowel Project B14 clinical trial [Paik et al., 2004]. The three risk categories exhibited statistically different survival at ten years, with 30% of recurrence (distant metastasis) in the high-risk group. These results suggest that ER-positive patients with high ONCOTYPE risk scores are not treated optimally with five years of tamoxifen therapy.

These gene signatures were developed with patients treated with adjuvant tamoxifen, i.e. treated after surgery. Therefore, the clinical outcome fitted by these prediction models was the survival of the patients or the appearance of metastases (survival analysis, see Section 2.3) instead of the response to the treatment itself (only available in the neoadjuvant setting, see Figure 1.4). This implies that these gene signatures predicted the natural history of the tumors (prognosis) and the response to tamoxifen (prediction) as well, with these two components being difficult to dissect.

Benefit of chemotherapy The study of gene expression profiles before and after treatment with chemotherapy is potentially informative in terms of biology and prediction model. Fisher et al. showed that the treatment of a tumor with chemotherapy before surgery (neaoadjuvant therapy, see Figure 1.4) does not adversely affect the survival of the patient, and it provides an in vivo assessment of response to chemotherapy [Fisher et al., 1998]. It is the ideal scenario to study the molecular changes and to identify candidate genes associated with drug response and resistance. The gene profiles derived prior to and after treatment have the potential to predict clinical outcomes with particular chemotherapy agents in individual patients. Two gene signature for chemotherapy response have been subsequently reported.

Ayers et al. identified a signature of 74 genes that discriminated between responders and non-responders to neoadjuvant chemotherapy (complete pathological response, denoted by pCR) in a cohort of 24 patients [Ayers et al., 2004]. No single clinical indicator or gene yielded sufficiently good performance or pCR prediction. However, the signature combining the expression of several genes yielded high specificity buy low sensitivity, on a small independent set of 18 patients.

Chang et al. identified a signature of 92 genes for chemotherapy resistance developed from 24 breast cancer patients [Chang et al., 2005]. In this study, tumor samples were classified as either sensitive or resistant to chemotherapy on the basis of the residual volume of the tumor at the end of treatment. The signature yielded high sensitivity and specificity

in cross-validation [Stone, 1974]. The investigators then validated their results using a small independent set of 6 patients, whereby all were correctly classified.

Both of these studies used datasets that were too small to draw statistically robust conclusions about the performance of such prediction models. However, they do support the concept that predictors of chemotherapy response can be developed.

1.4 Translational Research

The concept of *translational research* has been a center of focus in the biomedical community over the last few years, being viewed as a new way of thinking about and conducting life sciences research to accelerate healthcare outcomes [Woolf, 2008].

Translational research transforms scientific discoveries arising from laboratory or clinical studies into clinical applications ("from bench to bedside") to reduce cancer incidence, morbidity, and mortality (Figure 1.5).

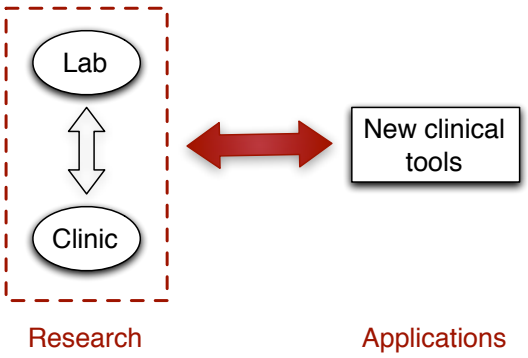


Figure 1.5: Translational research. Discoveries arising from laboratory or clinical studies are *translated* into new clinical tools. The red double arrow represents such a translation.

By removing barriers to interdisciplinary collaboration, translational research has the potential to drive the advancement of molecular-based medicine. By enabling doctors to leverage high throughput technologies, translational research could provide efficient clinical tools for early detection of cancer and other diseases, for improving drug development, and for making personalized medicine possible.

The movement of scientific discoveries into the clinic will accelerate only once doctors, biologists, bioinformaticians and the various operational members of staff can work together efficiently. To achieve this goal, translational research requires researchers and clinicians to have ready access to two critical types of information: (i) clinical information, including data contained in hospital systems and medical records, and pathology reports; and (ii) biomolecular information, including genomics, proteomics, medical imaging and other high throughput data.

The analysis of such data is one of the objectives of the Functional Genomics Unit at the Institut Jules Bordet in Brussels, headed by Prof. Christos Sotiriou, acting in liaison between the basic research laboratory and the clinical research setting to ensure faster application of experimental findings to the clinic. We will see in Section 6.2.4 that some of the findings presented in this thesis have been patented and commercialized in order to be used in day-to-day clinical practice.

1.5 Bioinformatics Context

In the early days of gene expression profiling studies, researchers used traditional statistics to analyze these new data. In the process, numerous problems arose around the issues of the small sample size, high dimensionality of the data, high level of noise and correlations of variables (gene co-expressions, see Section 2.1.2), rendering traditional methods unsuccessful [Zupan et al., 1999].

At that stage, machine learning techniques were picked up as candidates to circumvent such difficulties. Indeed, machine learning, defined as a field of artificial intelligence related to data mining and statistics, involves learning from data by dealing specifically with the curse of dimensionality (small sample size compared to the dimensionality of the data, also called the *high feature-to-sample ratio*), and the presence of noise [Mitchell, 1997].

During the early stages of microarray analysis, researchers used simple to complex machine learning techniques with some success, leading to high impact publications ([Golub et al., 1999; Brown et al., 2000; Alizadeh et al., 2000; van't Veer et al., 2002; Ramaswamy et al., 2003] to name a few). However, the pitfalls in the analysis of microarray data for classification tasks were emphasized by many authors [Simon, 2003; Simon et al., 2003; Michiels et al., 2005]. This includes the risk of overfitting [Everitt, 2002; Hastie et al., 2001] due to the high feature-to-sample ratio and the lack of validation due to the absence of independent datasets or incorrect use of cross-validation techniques [Stone, 1974]. Recently, Dupuy and Simon reviewed the statistical methods used in 90 microarray studies published before 2005 [Dupuy and Simon, 2007]. The authors extended the previous reviews from classification to class discovery and feature selection (called "outcome-related gene finding" in the article), drawing guidelines on statistical analysis. They found that 15% to 80% of the articles have at least one major flaw, depending on the impact of the journal (see supplementary information of [Dupuy and Simon, 2007] for a detailed list of flaws).

As the field became increasingly mature, it was unclear which methods yielded good performance in microarray analysis. To address this issue, several authors conducted large comparison studies of classification methods for microarray data [Ben-Dor et al., 2000; Dudoit et al., 2002]. They found that simple methods (e.g. linear models) yielded similar results or even outperformed complex ones (e.g. artificial neural networks, classification trees or support vector machines) in several microarray datasets. These results, challenging the use of complex classification models instead of simple ones, fundamentally changed the practice of classification analysis of microarray data.

At the time work was begun on this thesis, few articles reported the use of methods from survival analysis (Section 2.3) to build prognostic models from gene expression data, especially in breast cancer [Zupan et al., 1999]. Moreover, in contrast to the field of classification described above, the field lacked large comparative studies of survival models. However, there was an increasing interest in survival analyses of microarray data, the hope being to improve current prognostic classification.

1.6 Justification of the Thesis

The medical context of this thesis has emphasized the need for improving prognostic and predictive models in order to spare breast cancer patients from unnecessary anti-cancer treatments (and their concomitant adverse side effects), as well as to predict the response

or resistance to such treatments. The aim of this thesis is the development of novel methods to extract from microarray and survival data, the relevant molecular markers able to improve traditional prognostic models, bringing at the same time new insights into breast cancer biology.

The complexity of microarray data (high feature-to-sample ratio, high level of noise, gene co-expressions) makes their analysis a challenging task. In addition to their intrinsic complexity, the analyst has to deal with dilemmas arising in real clinical studies, such as the heterogeneity of populations of breast cancer patients under study and the use of different microarray technologies to carry out the gene expression profiling. When this thesis was begun, the field lacked the robust methods necessary to address these issues efficiently.

From a bioinformatics point of view, intense research efforts have been put into clustering and classification analyses of microarray data, leaving aside the analysis of survival data, which is particularly appealing for the development of prognostic models. Furthermore, numerous review articles have highlighted major flaws in the initial publications, such as the incorrect use of cross-validation techniques and the lack of validation data, shedding new light on these promising early results. It is therefore the right time to develop a new methodology dealing specifically with the microarray and survival data in order to derive an honest estimation of performance.

In this thesis, we will use machine learning techniques such as feature transformation/selection, local/additive prediction models, validation techniques and clustering, to develop such a methodology. We will show that the use of simple models in combination with *a priori* biomedical knowledge enables the building of efficient prognostic and predictive tools for breast cancer. Indeed, given the complexity of microarray and survival data, we will challenge the use of data-driven and complex models and demonstrate that simple models outperform complex ones in the framework of breast cancer prognostication. This is made possible by collecting numerous publicly available microarray datasets, which in turn enable us to conduct robust performance assessments and comparisons of the prognostic models.

Lastly, we will contribute to the field of breast cancer prognostication by implementing our novel methods in a publicly available R package and by identifying new prognostic gene signatures, which are the subject of high impact publications. The detailed list of the contributions of this thesis is provided below.

1.6.1 Contributions

We have introduced several new methods, which constitute key steps in the design of microarray survival analyses and yield innovative gene signatures and risk prediction models. In this section, we briefly describe these original contributions and we point the reader to the sections where the contribution has been introduced.

1.6.1.1 Methodological Contributions

Critical analysis of a clinical study using microarray data: We wrote a book chapter about a critical analysis of knowledge extraction from microarray and survival data in a clinical study [Haibe-Kains et al., 2008a]. The chapter describes each step of the data analysis procedure, from the quality control of data to the final validation, going through normalization, feature transformation, feature selection, and model building. Each section

proposes a set of guidelines and motivates the specific choice made for this particular clinical study.

Biology-driven feature transformation: We introduced a genome-wide and a prototype-based approach to reduce the dimensionality of the data, retaining relevant biological information and keeping the data interpretable. The genome-wide approach, detailed in Section 4.1.1, used a clustering approach in combination with gene annotations on an independent dataset to compute cluster centroids in the dataset under study. This approach was used in [Haibe-Kains et al., 2008a; Loi et al., 2008]. The second approach, detailed in Section 4.2.1, used *a priori* biological information to build modules of genes specifically correlated to a prototype, i.e. a key gene involved in a biological process of interest. This approach was used in [Desmedt et al., 2008].

Stability-based feature selection: We introduced a new statistic (Stab) to identify the signature size, i.e. the number of relevant features used to build a survival model, in the framework of feature ranking. This statistic, detailed in Section 4.1.2.2, assesses the stability of a signature of size k using a resampling procedure and allows for tuning the signature size without optimizing directly the model performance. This statistic was used in [Loi et al., 2008; Haibe-Kains et al., 2008a,c, 2009].

Robust model building: We used a robust model building approach consisting in the linear combination of univariate survival models (Section 4.1.3.1). Although such a procedure was well known in classification modeling, its application to survival analysis of microarray data in combination with stability-based feature selection have interesting properties and yielded good performance [Loi et al., 2008; Haibe-Kains et al., 2008c].

Modular modeling: The modular modeling approach, as detailed in Section 4.3.1, is based on a divide-and-conquer strategy that consists in attacking the prognostication of the global population of tumors by dividing it into subtypes, of which the specific survival models can be combined to yield a global model for prognostication. This method was used in [Haibe-Kains et al., 2009].

Performance assessment and comparison: We introduced a statistical framework to assess and to compare the performance of survival models. This framework, detailed in Section 4.4, is based on traditional statistics from survival analysis (but underused in microarray survival analysis) and (non-)parametric paired statistical tests. This framework was used in [Haibe-Kains et al., 2008c,b, 2009].

1.6.1.2 Software Contributions

Performance assessment and comparison: An `R` package, called `survcomp`, was implimented for the performance assessment and comparison of survival models. The package is fully documented and is publicly available from the comprehensive R archive network [CRAN].

SWEAVE code: The analysis for [Haibe-Kains et al., 2008a,c, 2009] complies with the *research reproducibility* guidelines proposed in [Gentleman, 2005], in terms of availability of the code and reproducibility of results and figures. All the codes are publicly available through `http://www.ulb.ac.be/di/map/bhaibeka/research.html`.

1.6.1.3 Biomedical Contributions

In addition to the new biological insights into breast cancer that we have introduced in our published articles, we have contributed to the field by identifying several prognostic gene signatures:

GGI: The gene expression grade index [Sotiriou et al., 2006] elucidated the molecular basis of histological grade, one of the most important clinical variables for breast cancer prognostication.

TAMR13: The tamoxifen resistance signature (TAMR13; [Loi et al., 2008]) was shown to be predictive of resistance to tamoxifen, the most widely used form of hormonotherapy in breast cancer.

Gene modules: The gene modules ESR1, ERBB2, AURKA, STAT1, VEGF, PLAU, CASP3 [Desmedt et al., 2008] allowed for a robust quantification of the key biological processes in breast cancer (Section 3.1). Using these gene modules, we revealed the thread connecting breast cancer molecular subtypes, prognostic gene signatures, and traditional clinico-pathological prognostic factors.

GENIUS: The gene expression prognostic index using subtypes [Haibe-Kains et al., 2009]) improved the state-of-the-art prognostic gene signatures by integrating the *a priori* biological knowledge about the existence of breast cancer molecular subtypes.

The list of annotated genes included in these signatures is provided in Appendix C.

Patents: A patent for the GGI signature was filled[2], while the patent for the tumor invasion and immune response modules have been submitted.

1.6.2 Publications

The articles published during the thesis are listed in the following sections. We distinguish between the papers related to this thesis and those out of the scope of this thesis (referred to as *collaborative papers*).

Thesis Related Papers

(Co-)First Author Note that the symbol * means that these authors contributed equally to the work.

- *Refining breast cancer prognostication according to the molecular subtypes.* **Haibe-Kains B***, Desmedt C*, Rothé F, Piccart MJ, Bontempi G and Sotiriou C, *submitted*, 2009.
- *Comparison of Prognostic Gene Expression signatures for Breast Cancer*. **Haibe-Kains B***, Desmedt C*, Piette F, Buyse M, Cardoso F, vant Veer L, Piccart MJ, Bontempi G and Sotiriou C in BMC Genomics, volume 9, number 394, September 2008.

[2] http://www.wipo.int/pctdb/en/wo.jsp?wo=2006119593

- *A comparative study of survival models for breast cancer prognostication based on microarray data: does a single gene beat them all?* **Haibe-Kains B**, Desmedt C, Sotiriou C and Bontempi G in Bioinformatics, volume 24, number 19, pages 2200-2208, July 2008.

- *Biological processes associated with breast cancer clinical outcome depend on the molecular subtypes*. Desmedt C*, **Haibe-Kains B***, Wirapati P, Buyse M, Larsimont D, Bontempi G, Delorenzi M, Piccart MJ, and Sotiriou C in Clinical Cancer Research, volume 14, number 16, pages 5158-5165, August 2008.

- *Predicting prognosis using molecular profiling in estrogen receptor-positive breast cancer treated with tamoxifen*. Loi S*, **Haibe-Kains B***, Desmedt C, Wirapati P, Lallemand F, Tutt AM, Gillet C, Ellis P, Ryder K, Reid JF, Daidone MG, Pierotti MA, Berns EMJJ, Jansen MPHM, Foekens JA, Delorenzi M, Bontempi G, Piccart MJ and Sotiriou C in BMC Genomics, volume 9, number 239, May 2008.

- *Definition of clinically distinct molecular subtypes in estrogen receptor positive breast carcinomas through use of genomic grade*. Loi S*, **Haibe-Kains B***, Desmedt C, Lallemand F, Tutt AM, Gillet C, Harris A, Bergh J, Foekens JA, Klijn J, Larsimont D, Buyse M Bontempi G, Delorenszi M, Piccart MJ and Sotiriou C in Journal of Clinical Oncology, volume 25, number 10, April 2007.

  - *Reply: Expression profiling in breast carcinoma: new insights on old prognostic factors?* Loi S, **Haibe-Kains B**, Desmedt C, and Sotiriou C in Journal of Clinical Oncology, volume 25, number 27, pages 4317-4318, September 2007.

Contributing Author

- *The Genomic Grade Index (GGI) Is Associated with Response to Neoadjuvant Chemotherapy in Patients with Breast Cancer*. Liedtke C, Hatzis C, Symmans WF, Desmedt C, **Haibe-Kains B**, Valero V, Kuerer H, Hortobagyi, Piccart MJ, Sotiriou C and Pusztai L in Journal of Clinical Oncology, *in press*, 2009.

- *Meta-analysis of Gene-Expression Profiles in Breast Cancer: Toward a Unified Understanding of Breast Cancer Sub-typing and Prognosis Signatures.* Wirapati P, Sotiriou C, Kunkel S, Farmer P, Pradervand S, **Haibe-Kains B**, Desmedt C, Ignatiadis M, Sengstag T, Schutz F, Goldstein DR, Piccart MJ and Delorenzi M in Breast Cancer Research, volume 10, number 4, R65, August 2008.

- *Strong time-dependency of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multi-centre independent validation series*. Desmedt C, Piette F, Loi S, Wang Y, Lallemand F, **Haibe-Kains B**, Viale G, Delorenzi M, Zhang Y, Saghatchian dAssignies M, Bergh J, Lidereau R, Ellis P, Harris A, Klijn JG, Foekens JA, Cardoso F, Piccart M, Buyse M and Sotiriou C in Journal of Clinical Cancer Research, volume 13, number 11, pages 3201-3214, June 2007.

- *Gene expression profiling in breast cancer: Understanding the molecular basis of histologic grade to improve prognosis*, Sotiriou C, Wirapati P, Loi S, Harris A, Bergh J, Smeds J, Farmer P, Praz V, **Haibe-Kains B**, Lallemand F, Buyse M, Piccart MJ and

Delorenzi M in Journal of National Cancer Institute, volume 98, pages 262-272, February 2006.

Book Chapter

- *Computational Intelligence in Clinical Oncology : Learned Lessons from a Case Study*. **Haibe-Kains B**, Desmedt C, Loi S, Delorenzi M, Sotiriou C, and Bontempi G, Chapter 10 in *Applications of Computational Intelligence in Bioinformatics and Biomedicine: Current Trends and Open Problems*, Springer-Verlag in the series *Studies in Computational Intelligence*, September 2008.

Collaborative Papers

Contributing Author

- *Gene expression profiling based on ZAP70 mRNA expression reveals differences in microenvironment interaction between patients with good and poor prognosis*. Stamatopoulos B, **Haibe-Kains B**, Equeter C, Meuleman N, Sorée A, De Bruyn C, Hanosset D, Bron D, Martiat P and Lagneaux L in Haematologica, *in press*, 2009.
- *MicroRNA-29c and microRNA-223 downregulation has in vivo significance in chronic lymphocytic leukemia and improves disease risk stratification*. Stamatopoulos B, Meuleman N, **Haibe-Kains B**, Saussoy P, Van Den Neste E, Michaux L, Heimann P, Martiat P, Bron D and Lagneaux L in Blood, *in press*, 2009.
- *Knocking Down Galectin 1 in Human Hs683 Glioblastoma Cells Impairs Both Angiogenesis and Endoplasmic Reticulum Stress Responses*. Le Mercier M, Mathieu V, **Haibe-Kains B**, Bontempi G, Mijatovic T, Decaestecker C, Kiss R and Lefranc F in Journal of Neuropathology and Experimental Neurology, volume 67, number 5, pages 456-469, May 2008.
- *Nucleolus and c-Myc: potential targets of cardenolide-mediated antitumor activity*. Mijatovic T, De Neve N, Gailly P, Mathieu V, **Haibe-Kains B**, Bontempi G, Lapeira J, Decaestecker C, Facchini V and Kiss R in Molecular Cancer Therapeutics, volume 7, number 5, pages 1285-1296, May 2008.
- *UNBS5162, a Novel Naphthalimide That Decreases CXCL Chemokine Expression in Experimental Prostate Cancers*. Mijatovic T, Mahieu T, Bruyère C, De Nève N, Dewelle J, Simon G, Dehoux MJM, van der Aar E, **Haibe-Kains B**, Bontempi G, Decaestecker C, Van Quaquebeke E, Darro F and Kiss R in Neoplasia, volume 10, number 6, pages 573-586, May 2008.
- *Evidence of galectin-1 involvement in glioma chemoresistance*. Le Mercier M, Lefranc F, Mijatovic T, Debeir O, **Haibe-Kains B**, Bontempi G, Decaestecker C, Kiss R and Mathieu V in Toxicology and Applied Pharmacology, volume 229, number 2, pages 172-183, January 2008.

- *Quantification of ZAP70 mRNA in B Cells by Real-Time PCR Is a Powerful Prognostic Factor in Chronic Lymphocytic Leukemia*. Stamatopoulos B, Meulemans N, **Haibe-Kains B**, Duvillier H, Massy M, Martiat P, Bron D, and Lagneaux L in Clinical Chemistry, volume 53, number 10, August 2007.

- *4-IBP: A s1 Receptor Agonist Decreases the Migration of Human Cancer Cells Including Glioblastoma Cells In Vitro and Sensitizes Them In Vitro and In Vivo to the Cytotoxic Insults of Pro-Apoptotic and Pro-Autophagic Drugs*. Mégalizzi V, Mathieu V, Mijatovic T, Gailly P, Debeir O, De Neve N, Van Damme M, Bontempi G, **Haibe-Kains B**, Decaestecker C, Kondo Y, Kiss R and Lefranc F in Neoplasia, volume 9, number 5, May 2007.

- *Gene regulation by phorbol 12-myristate 13-acetate in MCF-7 and MDA-MB-231, two breast cancer cell lines exhibiting highly different phenotypes*, Lacroix M, **Haibe-Kains B**, Hennuy B, Laes JF, Lallemand F, Gonze I, Cardoso F, Piccart MJ, Leclercq G and Sotiriou C in Oncology Reports, volume 12, number 4, pages 701-708, October 2004.

1.7 Glossary

Adjuvant! Online The goal is to help health professionals make estimates of the risk of poor outcome (cancer related mortality or relapse) without systemic adjuvant therapy, estimates of the reduction of these risks afforded by therapy, and risks of side effects of the therapy. These estimates are based on information entered about individual patients and their tumors (e.g. patient age, tumor size, nodal involvement or histological grade) These estimates are then provided on printed sheets in simple graphical and text formats to be used in consultations.

Adjuvant therapy Therapy given after the breast surgery to increase the chances of a cure. Adjuvant therapy may include chemotherapy or hormonotherapy.

Comparative genomic hybridization Comparative genomic hybridization (CGH) is a molecular-cytogenetic method for the analysis of copy number changes (gains/losses) in the DNA content of a given subject's DNA and often in tumor cells. The method is based on the hybridization of fluorescently labeled tumor DNA and normal DNA to to normal metaphase chromosomes (or DNA probes). Using epifluorescence microscopy and quantitative image analysis, regional differences in the fluorescence ratio of gains/losses vs control DNA can be detected and used for identifying abnormal regions in the genome. CGH will detect only unbalanced chromosomes changes. Structural chromosome aberrations such as balanced reciprocal translocations or inversions can not be detected since they do not change the copy number.

Cross-hybridization The hydrogen bonding of a single-stranded DNA sequence (see *hybridization*) that is partially but not entirely complementary to a single-stranded substrate. For instance, this can involve 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 sub-

sets 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 dissimilarity from the two elements or the two groups that it connects.

Distant metastasis Type of recurrence (see *Recurrence*). Tumor initiated from the primary breast tumor cells and that is located in another organ.

Estrogen receptor Estrogen receptor refers to a group of receptors which are activated by the hormone 17β-estradiol (estrogen; [Dahlman-Wright et al., 2006 Dec]). The estrogen receptor (ER) is a member of the nuclear hormone family of intracellular receptors. The main function of the ER is as a DNA binding transcription factor which regulates gene expression. However the ER also has additional functions independent of DNA binding.

Expressed sequence tag An expressed sequence tag or EST is a short sub-sequence of a transcribed cDNA sequence [Adams et al., 1991]. They may be used to identify gene transcripts, and are instrumental in gene discovery and gene sequence determination. ESTs represent portions of expressed genes. The current understanding of the human set of genes includes the existence of thousands of genes based solely on EST evidence. ESTs contain enough information to permit the design of precise probes for DNA microarrays that then can be used to determine the gene expression. Some authors use the term ”EST” to describe genes for which little or no further information exists besides the tag.

Gene expression Concentration of messenger RNA (mRNA) after transcription of a gene.

Gene Ontology The *Gene Ontology* project, or GO, provides a controlled vocabulary to describe gene and gene product attributes in any organism. It can be broadly split into two parts. The first is the ontology itself, actually three ontologies, each representing a key concept in molecular biology: the molecular function of gene products; their role in multi-step biological processes; and their localization to cellular components. The second part is annotation, the characterization of gene products using terms from the ontology. The members of the GO Consortium submit their data and it is made publicly available through the GO website[3].

Genotype The entire set of genes (genetic constitution) of an organism.

Histological Relating to *histology*.

Histology Study of the microscopic anatomy of cells and tissues. It is performed by examining a thin slice (section) of tissue under a light microscope or electron microscope. The ability to visualize or differentially identify microscopic structures is frequently enhanced through the use of histological stains.

Histo-pathological Relating to *histo-pathology*.

[3] `http://www.geneontology.org/`

Histo-pathology Microscopic study of diseased tissue, is an important tool in anatomical pathology, since accurate diagnosis of cancer and other diseases usually requires histopathological examination of samples. Trained medical doctors, namely pathologists, are the scientists who perform histopathological examination and provide diagnostic information based on their observations.

Human Epidermal growth factor Receptor 2 HER2 (also known as HER2/neu, ErbB-2 or ERBB2) stands for "Human Epidermal growth factor Receptor 2" and is a protein giving higher aggressiveness in breast cancers. It is a member of the ErbB protein family, more commonly known as the epidermal growth factor receptor family.

Hybridization Nucleic acid hybridization is the process of binding two complementary strands of DNA. 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.

Inflammatory response Inflammation occurs when tissues are injured by viruses, bacteria, trauma, chemicals, heat, cold or any other harmful stimulus. Chemicals including bradykinin, histamine, serotonin and others are released by specialised cells. These chemicals attract tissue macrophages and white blood cells to localise in an area to engulf (phagocytize) and destroy foreign substances. A byproduct of this activity is the formation of pus, a combination of white blood cells, bacteria and foreign debris.

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.

Malignancy Cancerous cells that usually have the ability to spread, invade, and destroy tissue. Malignant cells tend to have fast, uncontrolled growth due to changes in their genetic makeup.

Mer Refers to the length of a probe sequence, e.g. 60-mer probe is a probe composed of 60 nucleotides (see *Microarray*).

Microarray A DNA microarray is a multiplex technology used in molecular biology and in medicine. It consists of an arrayed series of thousands of microscopic spots of DNA oligonucleotides, called features, each containing picomoles of a specific DNA sequence. This can be a short section of a gene or other DNA element that are used as probes to hybridize a cDNA or cRNA sample (called target) under high-stringency conditions. Probe-target hybridization is usually detected and quantified by fluorescence-based detection of fluorophore-labeled targets to determine relative abundance of nucleic acid sequences in the target.

Neoadjuvant Therapy given as a first step to shrink a tumor before the breast surgery. Examples of neoadjuvant therapy include chemotherapy, or hormone therapy.

Nottingham Prognostic Index An index based on a combination of three prognostic factors: tumour size (cm x 0.2); lymph node stage (1 if node negative, 2 if ≤ 3 metastatic nodes, 3 if > 3 metastatic nodes) and histological grade (1, 2, or 3, for lo, intermediate and high histological grade respectively). Alternatively, lymph nodes can be classified according to level of involvement. A prognostic index < 3.4 implies a good prognosis, in $[3.4, 5.4]$ a moderately good prognosis and > 5.4 a poor prognosis.

Pathological Indicative of disease through examination of organs or tissues for instance. Relating to *pathology*.

Pathological complete response A pathological complete response is defined as the complete disappearance of tumor cells in the breast and the axillary lymph nodes after a neoadjuvant therapy.

Pathology Study and diagnosis of disease through examination of organs, tissues, bodily fluids and whole bodies.

Pathway Biological network that relates to a specific physiological process or phenotype. Set of linked biological components interacting with each other over time to generate a single biological effect

Personalized Medicine Currently, much of medical practice is based on ”standards of care” that are determined by averaging responses across large cohorts. The theory has been that everyone should get the same care based on clinical trials. *Personalized Medicine* is the concept that managing a patient's health should be based on the individual patient's specific characteristics, including age, gender, height/weight, diet, environment, or genetic profile using the high throughput technologies described in this thesis (Section 2.1).

Phenotype The expression of a particular trait, for example, skin color, height, behavior, according to the individuals genotype and environment.

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

Probes See *Microarray*.

Prognostication Prediction of the prospect of remission from the usual course of disease. In other words, prognostication refers to the prediction of the clinical outcome of a cancer patients in absence of anti-cancer treatment.

Prospective The two observation plans for clinical studies are the prospective and the retrospective ones. For the retrospective observation plan, the survival data are retrieved from patients medical histories. 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, with the time at which the events of interest occur being recorded. However, this observation plan is difficult to set up in practice, in that the investigator has to wait the end of follow-up before getting the final survival data.

Prediction Prediction of the response or resistance of a patient to an anti-cancer treatment. In other words, prediction refers to the prediction of the clinical outcome of a patient when an anti-cancer treatment is prescribed (see *adjuvant* and *neoadjuvant*).

Recurrence Occasionally breast cancer can return after primary treatment.

Relapse See *Recurrence*.

Retrospective The two observation plans for clinical studies are the prospective and the retrospective ones. For the retrospective observation plan, the survival data are retrieved from patients medical histories.

Reverse Transcriptase Polymerase Chain Reaction 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 > 1 is to stretch the PDF. The greater the magnitude, the greater the stretching. The effect of a scale parameter < 1 is to compress the PDF. The compressing approaches a spike as the scale parameter goes to zero. A scale parameter = 1 leaves the PDF unchanged (if the scale parameter is 1 to begin with) and non-positive scale parameters are not allowed.

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.

Single nucleotide polymorphism A single-nucleotide polymorphism (SNP, pronounced *snip*) is a DNA sequence variation occurring when a single nucleotide – A, T, C, or G – in the genome (or other shared sequence) differs between members of a species (or between paired chromosomes in an individual). For example, two sequenced DNA fragments from different individuals, AAGC**C**TA to AAGC**T**TA, contain a difference in a single nucleotide. In this case we say that there are two alleles : C and T. Almost all common SNPs have only two alleles. Variations in the DNA sequences of humans can affect how humans develop diseases and respond to pathogens, chemicals, drugs, vaccines, and other agents. SNPs are also thought to be key enablers in realizing the concept of personalized medicine [Engle et al., 2006]. However, their greatest importance in biomedical research is for comparing regions of the genome between cohorts (such as with matched cohorts with and without a disease).

Systemic Treatment using substances that travel through the bloodstream, reaching and affecting cells all over the body.

Tamoxifen Tamoxifen (Nolvadex) is a drug, taken orally as a tablet, which interferes with the activity of estrogen, a female hormone. Estrogen can promote the development of cancer in the breast. Tamoxifen is approved by the U.S. Food and Drug Administration (FDA) for the prevention of breast cancer and for the treatment of breast cancer, as well as other types of cancer.

Translational Research Translational research transforms scientific discoveries arising from laboratory, clinical, or population studies into clinical applications to reduce cancer incidence, morbidity, and mortality.

1.8 Abbreviations

AFT Accelerated Failure Time.

AUC Area Under the Curve.

BIC Bayesian Information criterion.

BSC Brier score.

CDF Cumulative Distribution Function.

CGH Comparative Genomic Hybridization.

***C*-index** Concordance index.

CRAN Comprehensive R Archive Network.

CVPL Cross-validated partial likelihood.

DMFS Distant Metastasis Free survival.

EASE Enrichment Analysis Systematic Explorer.

ER Estrogen Receptor.

EST Expressed Sequence Tag.

FISH Fluorescent in situ hybridization.

GENE70 Gene signature introduced in [van't Veer et al., 2002].

GENE76 Gene signature introduced in [Wang et al., 2005].

GENIUS Gene Expression progNostic Index Using Subtypes [Haibe-Kains et al., 2009].

GG Gene expression Grade [Sotiriou et al., 2006].

GGI Gene expression Grade Index [Sotiriou et al., 2006].

GO Gene Ontology.

HER2 Human Epidermal growth factor Receptor 2.

HG Histological Grade.

IAUC Integrated area under the curve.

IBSC Integrated Brier score.

IGS Invasiveness Gene Signature [Liu et al., 2007]

IPA Ingenuity Pathway Analysis [Ingenuity Systems].

IHC ImmunoHistoChemistry.

i.i.d. Independent and identically distributed.

KM Kaplan-Meier.

LMN Local Model Networks.

LOOCV Leave-One-Out Cross-Validation.

NA Not Available.

ONCOTYPE Gene signature for tamoxifen resistance [Paik et al., 2004].

OS Overall Survival.

P53 Gene signature for P53 mutation [Miller et al., 2005].

PCR Polymerase Chain Reaction.

pCR pathological Complete Response.

PDF Probability density function.

PgR Progesterone Receptor.

PL Partial likelihood.

PRESS Predicted REsidual Sums of Squares [Allen, 1974].

RFS Relapse Free Survival.

ROC Receiver operating characteristic.

RS Recurrence Score [Paik et al., 2004].

RTPCR Reverse Transcription Polymerase Chain Reaction.

SE Sensitivity.

SNP Single Nucleotide Polymorphism.

SP Specificity.

WOUND Gene signature of fibroblast serum response [Chang et al., 2004]

1.9 Notations

n	Number of samples or patients.
p	Number of input variables (probes or genes).
p'	Number of features.
$X, Y, ...$	Upper case letters represent matrices.
$x, y, ...$	Lower case letters represent the realization of random variables or vectors.
$\mathbf{X}, \mathbf{x}, ...$	Bold letters represent random variables.

X	Set of input variables, i.e. gene expressions.
X'	Set of features.
B	Set of objects (patients or genes).
K	Set of clusters, i.e. set of objects b.
g	Indicator variable for class ($g = 0$ for the low-risk class and $g = 1$ for the high-risk class).
$\mathcal{S}$	Scoring function.
t_i	Time of event occurrence for the sample i, $i \in \{1, ..., n\}$.
δ_i	Censoring indicator variable for the sample i, $i \in \{1, ..., n\}$.
β	Coefficients of a linear regression model.
$\hat{\beta}$	Estimated coefficients.
$S(t)$	Survivor function depending on time t.
$h(t)$	Hazard function depending on time t.
$S(t)$	Survival function depending on time t.
se	Standard error.
θ	Set of parameters.
ν	Number of parameters.
L	Likelihood function.
PL	Partial likelihood function.
$CVPL$	Cross-validated partial likelihood function.
E	Expectation of a random variable.
E^*	Expectation of a random variable from the uniform distribution.
$C(.)$	Clustering function.
$d(.,.)$	Dissimilarity (distance) function.
$W(.)$	Within cluster dissimilarity function.
$Gap(.)$	Gap function.
$\mathcal{N}$	Probability of Normal (Gaussian) distribution.
$BIC(.)$	Bayesian information criterion function.
D	Co-membership matrix.
ps	Prediction strength function.
$SE(.,.,.)$	Sensitivity function.
$SP(.,.,.)$	Specificity function.
$ROC(t)$	Receiver operating characteristic function depending on time t.
$BSC(t)$	Brier score function depending on time t.
$PRESS(m)$	Error in LOOCV of a linear model m as computed by the PRESS statitstic [Allen, 1974].

Chapter 2

Preliminaries

This chapter details the data and the state-of-the-art methods used in this thesis. The outline of the chapter is the following. First, we describe in Section 2.1 the microarray technology in terms of existing platforms and data generated. The characteristics of these data and the traditional approach used to analyze such data are then introduced. Second, we present the clustering methods, more particularly the hierarchical (Section 2.2.1) and model-based (Section 2.2.2) clusterings. Lastly, survival analysis, set of key methods for breast cancer prognostication and prediction, is presented in detail in Section 2.3.

2.1 Microarray Technology

To understand microarray technology, it is mandatory to have insight into the central dogma of molecular biology [Crick, 1970], namely the production of proteins from DNA as illustrated in Figure 2.1. Briefly, a specific sequence of DNA, called a gene, is translated into pre-mRNA by the means of RNA polymerase. This RNA is then usually modified (splicing) by an RNA-protein complex called the spliceosome[1]. Once the pre-mRNA is processed (maturation), the resulting mRNA message is then translated by the ribosome in order to produce proteins (translation). The expression of a particular gene is defined as the level (density) of mRNA produced by the transcription of this gene. The set of all gene expressions is called the transcriptome.

Microarray technology makes it possible to conduct expression profiling of thousand of genes in parallel. The concept behind this technology relies on accurate binding, also called *hybridization*, of strands of DNA with their precise complementary copies in experimental conditions where one sequence is also bound onto a solid state substrate (glass). Basically, a microarray chip is composed of DNA fragments (probes) that represent specific gene coding regions (see Figure 2.2). Purified RNA fragments from a biological sample are then fluorescently or radioactively labeled and hybridized to the chip (see Figure 2.3). Once the hybridization is complete, the chip is washed to remove non-hybridized fragments. The chip is then processed by a laser scanner in order to detect the areas of the chip where hybridizations occurred (see Figure 2.4).

[1] Sometimes a pre-mRNA message may be spliced in several different ways, allowing a single gene to encode multiple proteins (alternative splicing).

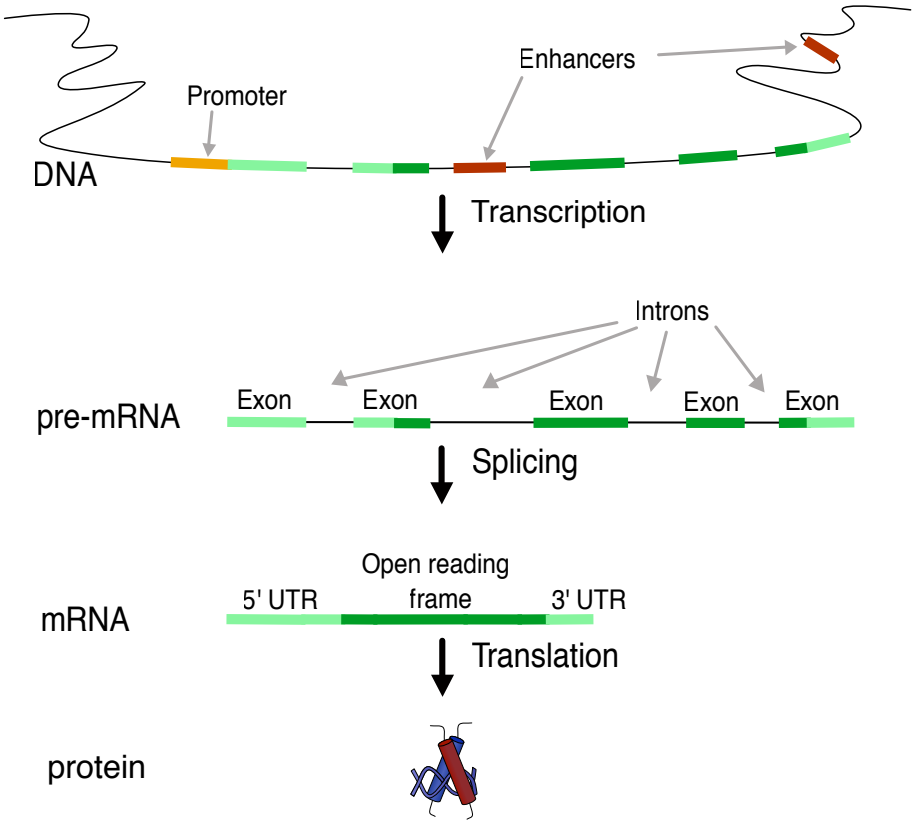


Figure 2.1: Biology dogma, from the DNA (gene) to the protein. Please refer to [Werner, 2005] for a detailed description of the central dogma of molecular biology. Image from [Wikipedia].

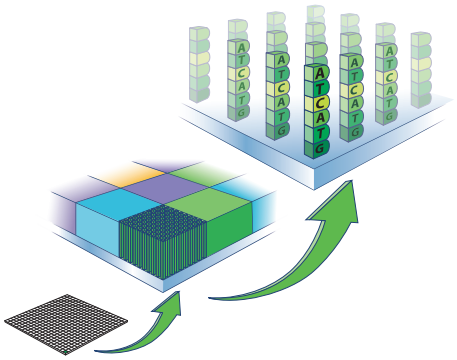


Figure 2.2: Different views of a microarray chip. Left: the whole chip. Middle: a zoom on a specific area of the chip with different set of probes. Right: a detailed view of each DNA strand. Image from [Affymetrix].

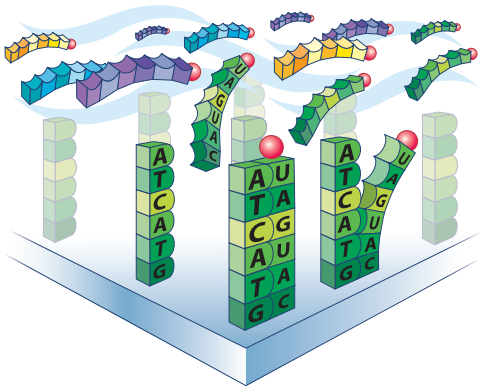


Figure 2.3: Hybridization of purified fragments of RNA obtained from a biological sample of interest. The fragments of RNA are labeled (red sphere). Image from [Affymetrix].

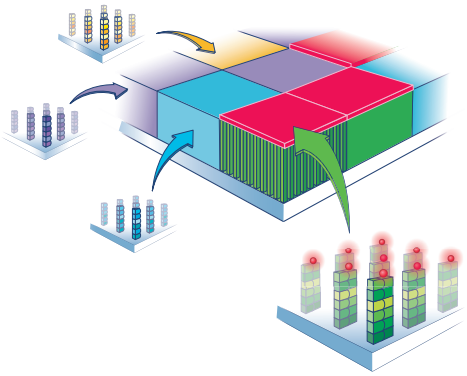


Figure 2.4: The labels (red spheres) are detected by the laser scanner in order to quantify the level of hybridization (if any) for each area of the chip. Image from [Affymetrix].

2.1.1 Microarray Platforms

Several variants of this technology have emerged since the late 1990s. The different microarray platforms can be classified with respect to their manufacturing (*spotted cDNA* or *oligonucleotide*) and hybridization quantification (*single* or *dual-channel*).

In spotted microarrays, the probes are synthesized prior to deposition on the array surface and are then ”spotted” onto the chip. A common approach utilizes an array of fine pins or needles controlled by a robotic arm that is dipped into wells containing DNA probes and then depositing each probe at specific locations on the array surface. This technique is mainly used by research scientists to produce ”in-house” microarrays since one can easily customize the probes and printing locations on the arrays. This provides a relatively low-cost microarray that may be customized for each study, and avoids the costs of purchasing often more expensive commercial arrays. However, Bammler et al. reported that in-house spotted microarrays may not provide the same level of sensitivity as commercial oligonucleotide arrays [Bammler et al., 2005].

In oligonucleotide microarrays, the probes are short sequences designed to match parts of the sequence of known or predicted gene coding regions. Oligonucleotide arrays are produced by printing short oligonucleotide sequences. These sequences are synthesized directly onto the array surface. Sequences may be longer (60-mer probes such as the Agilent design) or shorter (25-mer probes produced by Affymetrix) depending on the desired purpose; longer probes are more specific to individual target genes, and shorter probes may be spotted in higher density across the array and are cheaper to manufacture. One technique used to produce oligonucleotide arrays includes photolithographic synthesis (Agilent and Affymetrix) on a silica substrate, where light and light-sensitive masking agents are used to build a sequence one nucleotide at a time across the entire array [Pease et al., 1994].

In single-channel microarrays, a single RNA source is hybridized on a chip and comparison of RNA levels between samples is made in-silico in a post-processing phase of the experiment.

In dual-channel microarrays, two RNA sources are used, each labeled differently. The second RNA source is usually either a common reference against which all samples in an experiment are compared to, or a sample coming from a tissue under an alternative condition (e.g. disease vs non-disease), which allows direct comparison of RNA levels.

The number of probes and their composition depend on the microarray platform. The design of probes is a complex task since the sensitivity and specificity of a probe depend on the length and the sequence itself (see [Kane et al., 2000] for oligonucleotode microarray platforms). Additionally, microarray may contain large portion of *expressed sequence tag* (EST), i.e. transcribed sequence from unknown genes. This makes challenging the interpretation of the results obtained from the analysis of such microarrays, since a large number of probes lack of biological annotations. We refer the reader to [Murphy, 2002; Miller et al., 2002] for a review of issues in microarray design.

Table 2.1 gives a list of some widespread microarray platforms used in cancer research.

2.1.2 Microarray Data

Microarray technology is complex from a biological and a technical point of view (see Sections 2.1 and 2.1.1 respectively), and microarray data have some characteristics that make their analysis challenging:

Companies	Manufacturing	Hybridization quantification	Most recent chip	Probes
Applied Biosystems	Spotted cDNA	Single-channel	Human Genome survey Microarray v2.0	32,878
Eppendorf	Spotted cDNA	Single-channel	DualChip Microarray	294
Agilent	Oligonucleotide	Dual-channel	Whole Human Genome Oligo Microarray, G4112A	43,931
Affymetrix	Oligonucletide	Single-channel	HG-U133 Plus 2.0	54,675

Table 2.1: Widespread microarray platforms.

- High Dimensionality: Microarray technology generates a huge amount of data since it allows for the measurement of the expression of thousands of genes (whole genome) in parallel.
- High level of noise: A microarray experiment requires numerous steps, ranging from the preparation of the biological sample to the final quantification of the gene expressions. The purity of the samples as well as the technical variability inherent to each biological experiment, dramatically influence the quality of the data generated by microarray technology. Although these data are biologically informative, they are usually noisy (see [Chudin et al., 2001] for Affymetrix platform).
- Correlated measurements: The thousands of gene expressions measured through a microarray experiment are not independent. Indeed, numerous gene expressions are influenced by other genes, directly or indirectly. These interactions are partly explained by the existence of biological pathways, i.e. networks of spatiotemporal interactions between biological components as the products of gene expressions (proteins). These biological pathways may involve only a few to several hundreds of genes, leading to a high correlation of their gene expressions (also called gene co-expression, see Figure 2.5). We refer the reader to [Viswanathan et al., 2008] for a review of the analyses of biological pathways.

These characteristics must be taken into account for the analysis of microarray data, as we will see in the following sections.

2.1.3 Microarray Data Analysis

The analysis of microarrays is a complex task requiring biological and statistical expertise, such as sketched in Figure 2.6. First, a biological question of interest must be defined. An experimental design is then set up to assess the type and number of experiments to be carried out in order to be able to answer this question [The Tumor Analysis Best Practices

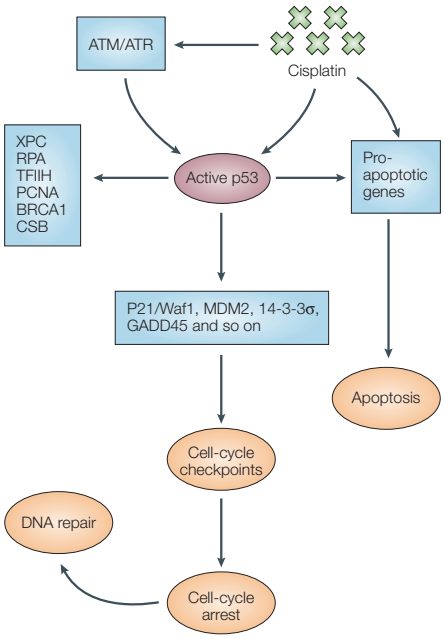


Figure 2.5: Example of biological pathway involving several genes relevant for breast cancer. Each directed arrow represents an interaction. Transduction of DNA-damage signals: p53 and cisplatin [Wang and Lippard, 2005].

Working Group, 2004; Affymetrix, 2004]. The microarray experiments are then effectively carried out to generate the microarray data. These data are preprocessed in order to control their quality and to remove systematic bias that may occur during experimentation. The data are then ready for analysis. Due to the huge number of gene expressions compared to the number of experiments (high feature-to-sample-ratio), it is often mandatory to reduce the dimensionality of the problem through feature transformation or selection. Depending on the biological question, different types of analyses are performed (e.g. unsupervised analysis like clustering, or supervised analysis like classification). Finally, the results are interpreted and validated. This may lead to new biological questions.

In this thesis, we focus on the feature transformation/selection and data analysis steps, while we use state-of-the-art methods for the data preprocessing step in our analyses. We briefly introduce these steps in the following sections.

2.1.3.1 Data Preprocessing

Profiling gene expressions is an expensive, time consuming and highly noisy process. As a consequence, it is essential to make the best use of the information contained in the gene expression data and to ascertain their quality. In this section, we will briefly describe quality controls and normalization procedures in case of Affymetrix technology.

Quality Controls Before starting the data analysis, preliminary checks are suggested in order to raise evidence of quality problems. In some cases, chips could appear beyond

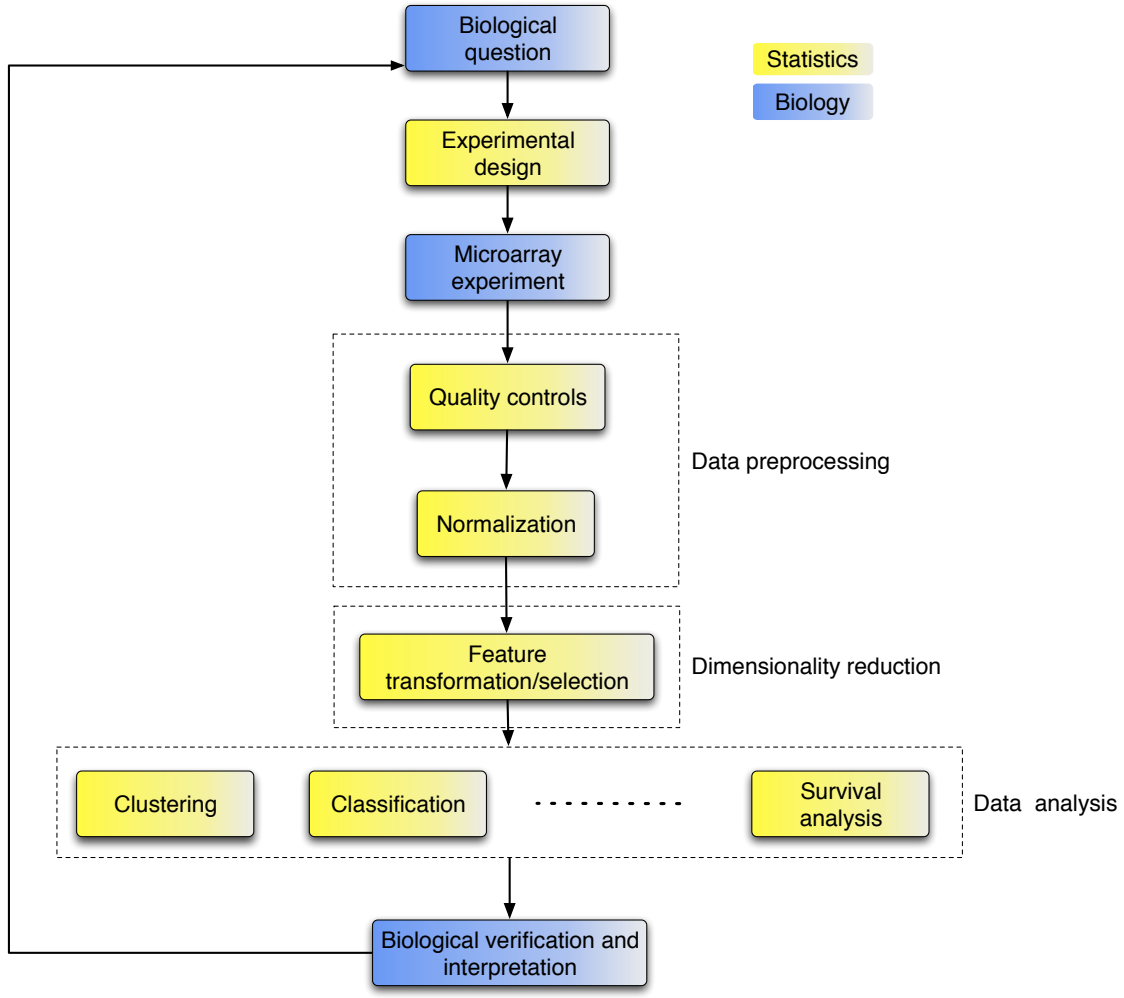


Figure 2.6: Overall design of an analysis of microarray data. Blue boxes refer to steps requiring biological expertise. Yellow boxes refer to steps requiring statistical analysis.

correction and the only recommended solution would be to discard them. For a review on existing methods for quality control we refer the reader to [Gentleman et al., 2005].

Here we will focus on the quality guidelines issued by [Affymetrix, 2002]. Two types of quality controls for Affymetrix chips are adopted :

- Single-chip quality controls: These controls concern one chip at a time. An example is the use of raw image analysis to detect hybridization artifacts like large areas of low intensity due to air bubbles.

- Multi-chip quality controls: These controls target a set of quantities whose "values should be comparable over all chips of a dataset" [Affymetrix, 2002], like scale factors, background intensities and percentage of present calls. Scale factors is a robust measure of the mean level of intensities on a chip. Background intensity is the intensity measured in an empty area (with no hybridization) and returns a measure of the background level. Percentage of present calls measures the proportion of genes being expressed (intensity significantly higher than background) on the chip. Once these quality controls have been carried out, the identification and the consequent discard of the anomalous chips is done.

Normalization Once the quality is assessed, an additional step is warranted to remove potential systematic bias, which may occur at each step of a microarray experiment (e.g. batch effect or scanner detection drift).

Normalization deals with systematic variations between experimental conditions (technical variation) which are not related to effective biological differences. Normalization methods aim to compensate for systematic technical differences between chips in order to enhance the analysis of biological differences between samples. Plenty of normalization methods specific to existing gene expression profiling technologies have been proposed in literature. Similarly to quality controls, they can be grouped in two main classes:

- Single-chip normalization methods: These are low complexity methods which use only one single chip to define the normalization transformation (e.g. mean scaling). A widely used single-chip normalization for Affymetrix technology is the Microarray Suite 5 (MAS5; [Affymetrix, 2002]).

- Multi-chip normalization methods: These methods use a set of chips to fit a (possibly) complex normalization transformation. This class of methods is sometimes referred to as *model-based* normalization methods. Widely used multi-chip normalization methods for Affymetrix technology are the Robust Multichip Average (RMA; [Irizarry et al., 2003]), RMA using sequence information (GCRMA; [Wu and Irizarry, 2004]), DNA-Chip Analyzer (dChip; [Li and Wong, 2001] and Variance Stabilization Normalization (VSN; [Huber et al., 2002]).

An overview of these normalization methods is given in [Gentleman et al., 2005]. Several studies addressed the question about the impact of normalization methods on gene expression analysis [Ploner et al., 2005; Bolstad et al., 2003; Harr and Schlotterer, 2006]. Although no gold standard has arisen in recent decades, guidelines have emerged from large comparison studies (see [Bolstad et al., 2003] for such a study for the Affymetrix platform). The

MicroArray Quality Control project[2] (MAQC), initiated by the Food and Drug Administration in the US, is expected to provide standards in the coming years.

The normalized microarray data for p genes and n patients are denoted by $X_{n\times p}$, such that x_{ij} represents the expression of gene j of patient i.

2.1.3.2 Dimensionality Reduction

There exist two main classes of methods to reduce the dimensionality of microarray data: feature transformation and feature selection. Depending of the design of the analysis, it can be decided that the analysis should use none of these methods (no dimensionality reduction), only one of them, or one method after the other (typically, feature transformation is performed first, followed by feature selection).

Feature transformation We refer to feature transformation as the method transforming the input space (genes) into a feature space without using outcome data (unsupervised method, see Figure 2.7 in the next section). The feature space is usually of a lower dimension in order to reduce the complexity of the data to analyze.

$$X_{n\times p} \rightarrow X'_{n\times p'} : p \gg p' \tag{2.1}$$

where X is the matrix of p gene expressions for n patient and X' is the matrix of p' features after transformation.

Some properties of feature transformation methods are recommended:

- Interpretability: If the final results of the analysis (e.g. a gene signature and its corresponding prediction model) have to be interpreted from a biological point of view, the features computed by the feature transformation methods have to be interpretable as well.
- Information: The new features should contain all the ”relevant” information from the original input space. The relevance of the new features depends on the biomedical question to address (outcome data). Since feature transformation methods do not use these supervised data, it is difficult to assess the relevance of the information after transformation before completing the whole analysis.
- Generalizability: Most methodologies do not assess the generalizability of feature transformation methods. This may lead to poor performance of the method in an independent dataset since a structure found in one dataset might not be generalizable to another dataset.

Three main methods are available for unsupervised feature transformation: compression, kernel and clustering methods. Compression and kernel methods transform the original input space into a new one, the dimensions of which are a linear combination of the original variables. These new variables (called features) are difficult to interpret from a biological point of view. Examples of compression and kernel methods are the principal component analysis [Jolliffe, 2002] and the kernel independent component analysis [Bach and Jordan,

[2] `http://edkb.fda.gov/MAQC/`

2003], respectively. Overview of compression and kernel methods are given in [Cristianini and Shawe-Taylor, 2000]. Clustering methods rely on the fact that many genes are co-expressed and that their expressions are highly correlated. The approach consists in finding clusters of highly correlated genes and in summarizing each set of clustered genes by the centroid (or prototype) of the cluster [Guyon and Elisseeff, 2003]. The transformed variables (called features) are expected to have lower variance than the original ones, yet remain easy to interpret from a biological and medical point of view.

Feature selection While the role of feature transformation is to reduce the dimensionality of the data without looking at the association with outcome data, feature selection [Kohavi and John, 1997; Guyon and Elisseeff, 2003] seeks which features, among the available ones, provide the largest amount of information for the prediction task (supervised method, see figure 2.8 in the next section). There are several benefits of feature selection: (i) facilitating data visualization and data understanding; (ii) reducing the measurement and storage requirements; (iii) reducing training and utilization times of the predictive model; and (iv) defying the curse of dimensionality to improve prediction accuracy.

Some properties of feature selection methods are recommended:

- Computational cost: Given the high dimensionality of microarray data, feature selection methods should be computationally effective.
- Information: The subset of selected features should yield good performance for the prediction task while keeping this subset small enough to enjoy the benefits of feature selection

There are three main categories of feature selection methods: filter, wrapper and embedded methods. Filter methods assess the relevance of features, ignoring the effects of the selected feature subset on the accuracy of the model. Wrapper methods assess subsets of features according to their relevance for a given model. These methods conduct a search for a good subset using the model itself as part of the evaluation function (e.g. forward, backward and stepwise feature selections). Embedded methods perform feature selection as part of the model fitting and are usually specific to given models (e.g. classification trees and regularization techniques).

In case of breast cancer prognostication, the predictive model is a survival model as described in Section 2.3.

2.1.3.3 Data Analysis

Depending on the biological question of interest, different analyses can be performed. There are two main classes of such analyses: unsupervised and supervised [Hastie et al., 2001; Webb, 2003].

An unsupervised analysis consists in finding a structure in the data without using external information (see Figure 2.7). The output of unsupervised analysis is usually a grouping of similar objects with respect to some criterion of similarity. This is referred to as clustering, a method that makes it possible to identify groups of patients with similar gene expressions (also called ”genetic portrait”) or groups of genes with similar expression (gene co-expression). Methods for clustering are described in detail in Section 2.2.

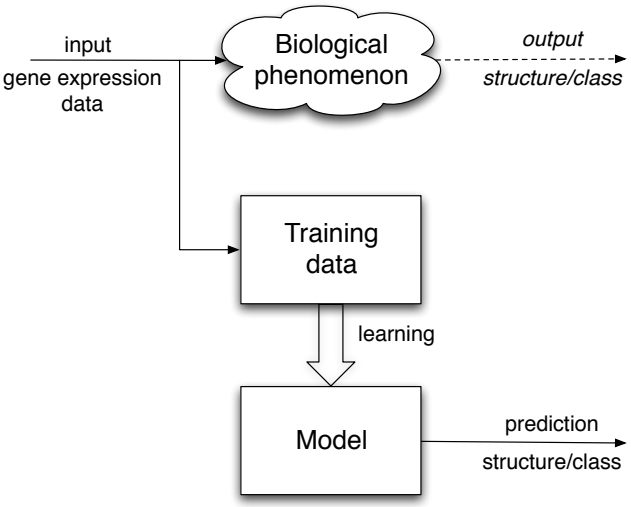


Figure 2.7: Unsupervised analysis. The output of the biological phenomenon (in *italic*) is hidden in the data and is not actually observed by the analyst.

As shown in Figure 2.8, a supervised analysis consists in finding a relationship (in the form of a statistical model) between input data (e.g. gene expressions) and output (e.g. response/resistance to a treatment or the survival of a patient). There are several types of supervised analysis, the choice of which depends on the output of interest: (i) classification analysis if the output is a discrete variable (binary or multi-class); (ii) regression analysis if the output is a continuous variable; or (iii) survival analysis if the output is survival data. Since survival analysis is intensively used in this thesis, it will be described in detail in Section 2.3.

2.2 Clustering

Cluster analysis is the grouping of objects (e.g. patients or genes) in a population in order to discover some structure in the data. The objects within a group should be similar to one another (i.e. share some traits to be defined by the analyst), but dissimilar from objects in other groups. Clustering is fundamentally a collection of methods of data exploration, often used to assess the presence of natural groupings in the data. If groups do emerge, their properties can be summarized to reduce the information on the original dataset to information about a small number of groups; alternatively, the original dataset can be divided to reduce the complexity of the problem. However, different methods yield different groupings, since each one implicitly imposes a structure on the data. Moreover, these techniques will produce groupings even if there is no ”natural” grouping in the data. Therefore, the analyst must be aware of the structures imposed by the methods and must choose them according to the problem under consideration.

In this section, we denote by B the set of $q \geq 2$ objects to be clustered. These objects are either the patients or the genes in microarray survival analysis but, for the sake of clarity, we will illustrate all the methods through the clustering of patients with respect to the expressions of few genes. Figure 2.9 illustrates such a cluster analysis of breast cancer patients for

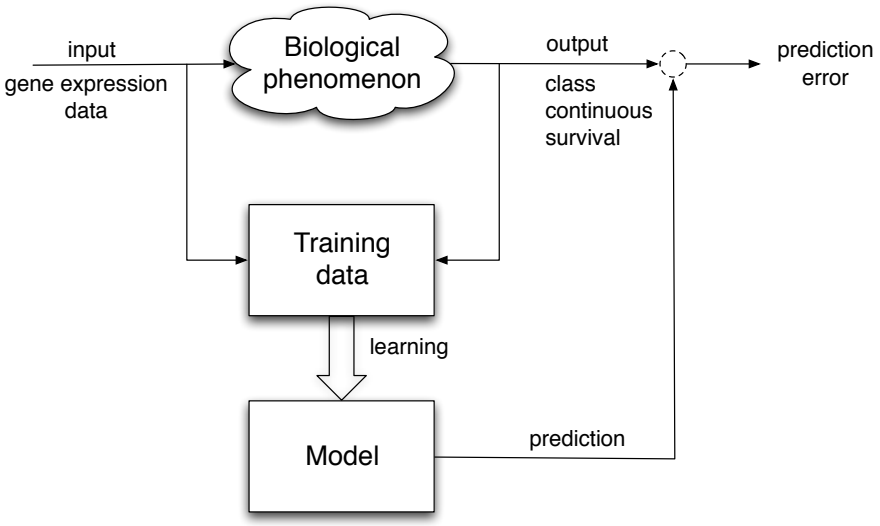


Figure 2.8: Supervised analysis. Since the output is actually observed, the prediction error of the model can be estimated (supervision).

whom we have measured the expression of two genes. Note that in biclustering [Cheng and Church, 2000; Sheng et al., 2003], the clustering is performed at both levels, i.e. patients and genes. In this thesis, we focus our research on hierarchical (Section 2.2.1) and model-based (Section 2.2.2) clusterings.

From a mathematical point of view, a clustering function $C(B)$ is a partitioning of the set B of $q \geq 2$ objects into a set K of u disjoint subsets (clusters) of objects with $1 \leq u \leq q$. $C(B)$ is defined as

$$C(B) : B \rightarrow K = \{k_1, \ldots, k_u\} : \qquad \bigcup_{j=1}^{u} k_j = B \tag{2.2}$$

$$\text{and} \quad i, j \in \{1, \ldots, u\}, i \neq j : k_i \cap k_j = \{\} \tag{2.3}$$

This definition is also referred to as *hard partitioning* since an object b is in a single cluster k_j. Relaxing conditions (2.3) to allow for an object to belong to one or more clusters leads to *soft partitioning* [Jain et al., 1999].

We mentioned above that the objects within a cluster should be similar to one another, but dissimilar from objects in other clusters. The definition of the (dis)similarity between objects is based on a notion of distance in the data space such as the Euclidean and correlation-based distances (see next section).

There is a vast corpus of literature on clustering and a wide range of application areas. Eisen et al. introduced this methodology in microarray studies [Eisen et al., 1998] but numerous clustering methods have been used since this seminal article: hierarchical clustering [Eisen et al., 1998; de Souto et al., 2008], k-means clustering [de Souto et al., 2008], partition around medoids [van der Laan et al., 2003], self-organizing maps [Tamayo et al., 1999], biclustering [Cheng and Church, 2000; Sheng et al., 2003] and quality-based clustering [De Smet et al., 2002; Tseng and Wong, 2005] to name a few. We will now describe hierarchical

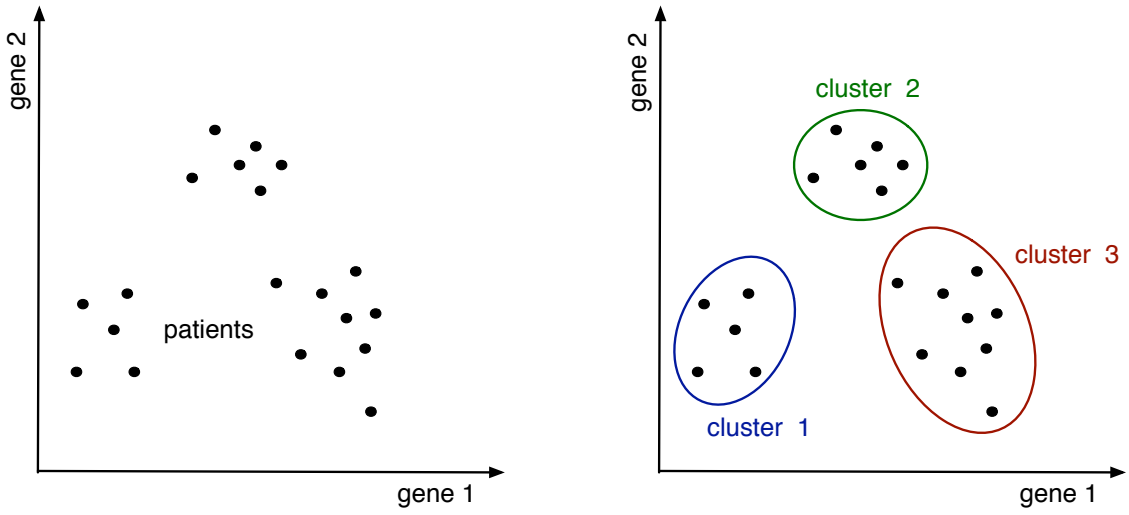


Figure 2.9: Example of clustering: (a) patients drawn in a two-dimensional space defined by the expression of two genes; (b) cluster analysis resulting in the discovery of three clusters.

clustering and mixture models, as these two methods are the most widely used clustering methods in microarray studies.

2.2.1 Hierarchical Clustering

Hierarchical clustering methods produce a hierarchical representation of the objects in the dataset. The clusters at each level of the hierarchy are created by merging clusters at the next lower level. At the lowest level, each cluster contains a single object. At the highest level, there is only one cluster containing all the objects. The height of each branch is proportional to the measure of dissimilarity between two merged clusters. Figure 2.10 illustrates such a hierarchical representation, also called *dendrogram*.

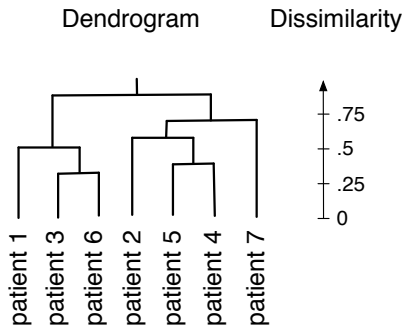


Figure 2.10: Example of hierarchical representation (dendrogram) produced by a hierarchical clustering analysis of 7 patients.

Hierarchical clustering methods are non-parametric, i.e. they do not rely on a probabilistic

model that generates the observed data. However, they require that the analyst specifies the strategy for building the dendrogram, the measure of dissimilarity (distance) and the linkage, i.e. the measure of dissimilarity between two clusters. We describe below the hierarchical clustering method using the *agglomerative* strategy with *correlation*-based dissimilarity and *average* linkage (see [Hastie et al., 2001; Webb, 2003] for a review of alternative strategies, dissimilarities and linkages).

Agglomerative building strategy Agglomerative building strategy is a bottom-up approach: the algorithm starts at the bottom of the dendrogram and at each level merges a selected pair of clusters into a single cluster. The pair chosen for merging consists of the two clusters with the smallest dissimilarity.

Correlation-based dissimilarity Let $d(b_i, b_j)$ be the dissimilarity function between objects b_i and b_j. The dissimilarity function used by hierarchical clustering methods is required to satisfy the following conditions:

- Non-negativity: $d(b_i, b_j) \geq 0 \ \forall i, j$
- Identification mark: $d(b_i, b_i) = 0 \ \forall i$
- Symmetry: $d(b_i, b_j) = d(b_j, b_i) \ \forall i, j$

Correlation-based dissimilarity defines dissimilarity between two objects b_i and b_j through the Pearson correlation coefficient ρ [Rodgers and Nicewander, 1988]. If highly correlated and anti-correlated objects have to be in the same cluster, the dissimilarity is defined as

$$d(b_i, b_j) = 1 - \left|\rho(b_i, b_j)\right| \tag{2.4}$$

In contrast, if highly anti-correlated objects have to be in different clusters, the dissimilarity is defined as

$$d(b_i, b_j) = 1 - \rho(b_i, b_j) \tag{2.5}$$

Average linkage The average linkage defines the dissimilarity between two clusters k_q and k_r as the average of dissimilarities between each pair of objects belonging to different clusters such that

$$d(k_r, k_s) = \begin{cases} \frac{1}{n_{k_r} n_{k_s}} \sum_{i \in k_r} \sum_{j \in k_s} d(b_i, b_j) & \text{if } r \neq s \\ 0 & \text{if } r = s \end{cases} \tag{2.6}$$

where n_{k_r} and n_{k_s} are the number of objects in clusters k_r and k_s respectively. Note that the case $r = s$ is never met in the procedure performing the hierachical clustering (Algorithm 1).

The algorithm building a hierarchical clustering using the agglomerative strategy, the dissimilarity and the linkage defined above is given in Algorithm 1. Let us illustrate this procedure with the hierarchical clustering of patients with respect to their gene expressions. Let X be the gene expression matrix for n patients, the set of clusters K being initialized with

one patient per cluster. The dendrogram *hcl* is initialized with this first set of clusters and a dissimilarity equal to 0. For each step of the while loop, the two least dissimilar clusters are identified. If the two clusters contain only the gene expression profile x of a single patient, the dissimilarity is estimated using Equation (2.4) or (2.5), otherwise, Equation (2.6) is used. Then the set of clusters K is updated by merging the two clusters of interest, and the dendrogram *hcl* is updated with the new set of clusters and the dissimilarity d'. Therefore *hcl* contains all the clusters to be merged and their dissimilarity at each step of the algorithm, enabling the construction of the dendrogram.

Algorithm 1 Hierarchical clustering

procedure HCLUST(X)
  $K \leftarrow \{x_1, ..., x_n\}$ ▷ each cluster contains a single object from X
  $hcl \leftarrow \{K, 0\}$ ▷ Initialization of the dendrogram *hcl* with K and dissimilarity 0
  while $|K| > 1$ **do** ▷ more than one cluster remained in K
    $\{i, j\} \leftarrow \underset{i,j}{\operatorname{argmin}} \left(d(k_i, k_j)\right), \forall i \neq j \in \{1, ..., |K|\}$ ▷ find the clusters i and j with the lowest dissimilarity
    $d' \leftarrow d(k_i, k_j)$
    $k_i \leftarrow k_i \cup k_j$ ▷ merge these two clusters in k_i
    $K \leftarrow K \setminus k_j$ ▷ remove cluster k_j
    $hcl \leftarrow \{hcl, \{K, d'\}\}$ ▷ update the dendrogram
  end while
  return *hcl*
end procedure

2.2.1.1 Number of Clusters

Although the hierarchical clustering method does not require the specification of u, the number of clusters, it is up to the analyst to cut the dendrogram at a particular level to produce u disjoint clusters leading to hard partitioning of the dataset (Figure 2.11). However, the analyst could cut the dendrogram at different levels, and an object could then belong to several clusters, rendering difficult the interpretation of the results[3]. This clustering should represent a "natural" grouping in the sense that objects within each cluster are sufficiently more similar to each other than to objects assigned to different clusters at that level. The Gap statistic [Tibshirani et al., 2001] can be used for the selection of u, the number of clusters.

The Gap statistic is a data-driven method to estimate u^*, the number of clusters present in the original dataset. This statistic is based on the "within cluster dissimilarity" W_u defined as

$$W_u(B) = \frac{1}{2} \sum_{r=1}^{u} \sum_{i \in k_r} \sum_{j \in k_r} d(b_i, b_j)$$

where the k_r are clusters resulting from the partitioning of the objects in u clusters.

W_u characterizes the extent to which objects assigned to the same cluster tend to be dissimilar to one another. The values $W_1, W_2, ..., W_u$ generally decrease with increasing u

[3] In the literature, the hierarchical clustering method is always used for hard partitioning.

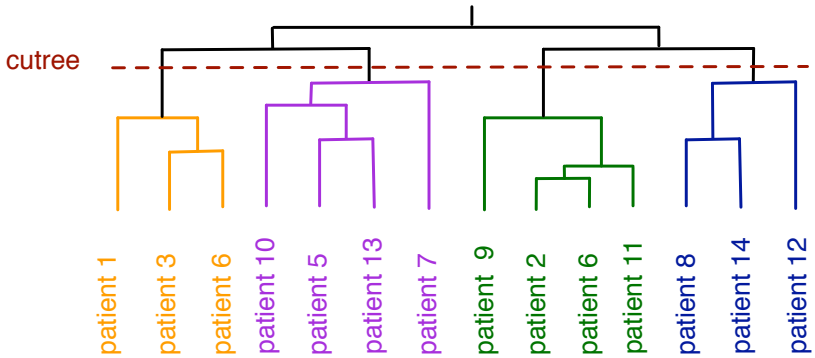


Figure 2.11: Example of dendrogram of 14 patients. The dendrogram is cut by the function *cutree* to get $u = 4$ clusters differentiated by colors.

since a large number of cluster centers will tend to fill the feature space densely, and thus will be closed to each object.

The intuition underlying the approach is that if there are actually u^* distinct clusters of objects, then for $u < u^*$ the clusters returned by the algorithm will each contain a subset of the true underlying clusters. That is, the clustering function will not assign objects in the same naturally occurring cluster to different estimated clusters. To the extent that this is the case, the within cluster dissimilarity value will tend to decrease substantially with each successive increase in the number of specified clusters, $W_{u+1} \ll W_u$, as the natural clusters are successively assigned to separate clusters. For $u > u^*$, one of the estimated clusters must partition at least one of the natural clusters into two subsets. This will tend to provide a smaller decrease in the criterion as u is further increased. Splitting a natural cluster, within which the objects are all quite similar to each other, reduces the criterion less than partitioning the union of two dissimilar ones into their proper constituents. So there will be a sharp decrease in $W_u - W_{u+1}$ at $u = u^*$. An estimate of u^* is then obtained by identifying a "kink" in the plot of W_u as a function of u. The Gap statistic compares the curve log W_u to the curve obtained from data uniformly distributed. It estimates the optimal number of clusters to be the place where the gap between these two curves is largest, i.e. maximizing

$$Gap_u(B) = E^*\{\log(W_u(B))\} - \log(W_u(B))$$

where E^* denotes the expectation from the uniform distribution.

2.2.2 Mixture Modeling

Mixture modeling assumes that the data is an Independent and identically-distributed (i.i.d.) sample from a population described by a probability density function. This density function is characterized by a parametrized model, taken to be a mixture of component density functions, where each component density describes one of the clusters. The population B of objects b is described by a finite mixture distribution of the form

$$\Pr(b) = \sum_{r=1}^{u} \pi_r \Pr(b|r)$$

where u is the number of clusters in the population, π_r are the mixing proportions such that $\sum_{r=1}^{u} \pi_r = 1$, and $\Pr(b|r)$ is the r^{th} probability density function of b. The quantity π_r is typically interpreted as the *prior* probability that a data point is generated by the r^{th} component of the mixture.

There are three sets of parameters to estimate: the values of π_r, the parameters of the probability distribution of each of the components, and the value of u. The usual approach to clustering using finite mixture distributions is first to specify the form of the component distributions, $\Pr(b|r)$. Then the number of clusters, u, is prescribed. The parameters of the model are estimated and the objects are grouped on the basis on their estimated *posterior* probabilities of cluster membership. Using Bayes' theorem, the object b is assigned to cluster r if

$$\Pr(r|b) \geq \Pr(s|b) \; \forall r \neq s \text{ with } r, s \in \{1, \ldots, u\}$$

$$\text{where} \qquad \Pr(r|b) = \frac{\pi_r \Pr(b|r)}{\sum_{s=1}^{u} \pi_s \Pr(b|s)} \tag{2.7}$$

Although this method leads to hard partitioning of the dataset, the analyst could easily use the probabilities of an object b to belong to each cluster to carry out soft partitioning.

The most widely used form of mixture distribution for continuous variables is the mixture of normal (Gaussians) distributions, where the r^{th} component $\Pr(b|r) \sim \mathcal{N}(\mu_r, \Sigma_r)$, where μ_r and Σ_r are the means and covariance matrix of a multivariate normal distribution. So

$$\Pr(b) = \sum_{r=1}^{u} \pi_r \mathcal{N}(b; \mu_r, \Sigma_r) \tag{2.8}$$

The estimation of the parameters of a normal mixture model can be achieved by the maximum likelihood procedure through the Expectation-Maximization (EM) algorithm [Dempster et al., 1977]. The convergence of the algorithm may be poor, depending on the data distribution. However, constraints on the covariance matrix (e.g. diagonal matrix) reduce the number of parameters to estimate, improving the convergence rate of the EM algorithm [Everitt and Hand, 1981; Celeux and Govaert, 1995].

2.2.2.1 Number of Clusters

The selection of u, i.e. the number of clusters, is not trivial (Section 2.2.1.1,). In the case of mixture models, this depends on many factors, for instance the shape of clusters, the sample size and the dimensions of the data.

The Bayesian information criterion (BIC) [Schwarz, 1978] can be used to estimate the likelihood of a mixture model with u clusters. The BIC is the value of the maximized log-likelihood with a penalty for the number of parameters in the model, and allows comparison of models with different parameterizations and/or different numbers of clusters. The BIC is defined as

$$BIC_u(B) = 2 \sum_{b \in B} \log \Pr(b; \theta_u) - \nu_u \log(|B|)$$

where B is the set of objects b, θ_u is the set of parameters of the mixture of u components, ν_u is the number of such parameters to be estimated (depending on the model), and $\sum_{b \in B} \log \Pr(b; \theta_u)$ is the likelihood of the data given the mixture model of parameters θ_u. The larger the value of the BIC, the stronger the evidence for the model and number of clusters [Fraley and Raftery, 1998; Yeung et al., 2001]. Therefore, we can select the number of cluster on the basis of the BIC.

2.2.3 Heatmap

A heatmap is a graphical representation of data, in which the values of the variables are represented as colors in a two-dimensional map. This makes is possible to visualize a large quantity of values, such as in microarray data. Although the heatmap is not part of the clustering algorithm itself, it represents an important step towards visualizing the results. Figure 2.12 illustrates the use of a heatmap in combination with hierarchical clustering in order to visualize the gene expressions.

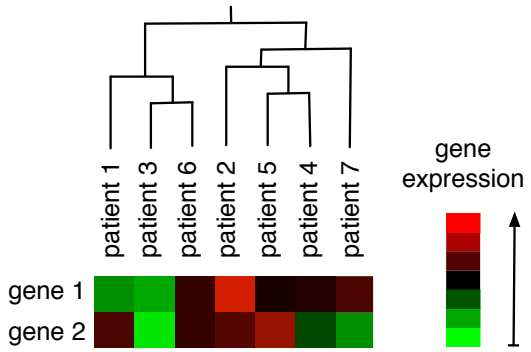


Figure 2.12: Example of heatmap in combination with a hierarchical representation (dendrogram) produced by a hierarchical clustering analysis of 7 patients for whom we measured the expression two genes.

Although most publications in microarray analysis include a heatmap in combination with hierarchical clustering, the use of heatmaps is not limited to this method. Once the clustering results (hard or soft partitioning) are generated, the objects can be sorted in order to reflect these results.

2.2.4 Performance Assessment

The performance assessment in clustering analysis is a difficult task since the ”truth” remains hidden to the analyst (unsupervised learning, see Figure 2.7). However numerous criteria have been designed in the past few years, most of them allowing for the assessment of the stability/variance of a clustering [Fraley and Raftery, 1998; Sugar, 1998; Tibshirani et al., 2001; Ben-Hur et al., 2002]. These criteria have also been used to identify the ”good” number of clusters, as described in Sections 2.2.1.1 and 2.2.2.1 for hierarchical and mixture modeling clustering, respectively. Recently, Tibshirani and Walther introduced a new framework for performance assessment of clustering analysis [Tibshirani and Walther, 2005]. The key idea

is to view clustering as a supervised classication problem, in which we must also estimate the true class labels. The resulting *prediction strength* measure assesses how well the clusters can be predicted from the data.

Let $X_{n\times p}$ and $Y_{m\times p}$ be two datasets of n and m objects to assign into u clusters in a p-dimensional space. Let $C_X(Y)$ denote the use, in the dataset Y, of the clustering model C fitted on the dataset X. To summarize the result of the clustering $C_X(Y)$, we define the co-membership matrix $D[C_X(Y)]_{m\times m}$ as

$$D[C_X(Y)]_{ii'} = \begin{cases} 1 \text{ if } i, i' \in k \\ 0 \text{ otherwise} \end{cases} \tag{2.9}$$

where $k \in K$ is a cluster of objects. In words, D is a matrix whose entry (i, i') is equal to 1 if the objects i and i' fall into the same cluster, 0 otherwise.

Considering X as a training set and Y as an independent validation set, the main idea is to: (i) cluster Y into u clusters ($C_Y(Y)$); (ii) cluster X into u clusters ($C_X(X)$); and (iii) measure how well C_X predicts co-membership in Y ($C_X(Y)$). We can define the prediction strength of the clustering function C as

$$ps = \min_{1\leq j\leq u} \frac{1}{n_{k_j}(n_{k_j}-1)} \sum_{i\neq i'\in k_j} D[C_X(Y)]_{ii'} \tag{2.10}$$

where the k_j's with $1 \leq j \leq u$, are the clusters defined by the clustering $C_Y(Y)$ and n_{k_j} is the number of objects in cluster k_j. In words, for each validation cluster defined by the cluster function trained on the validation set, we compute the proportion of object pairs in that cluster that are also assigned to the same cluster by the cluster function trained on the training set. The prediction strength is the minimum of this quantity over the u clusters in the validation set and lies in $[0, 1]$. Regarding the results obtained from simulated and real data experiments, the authors considered as good a prediction strength $ps \geq 0.8$ [Tibshirani and Walther, 2005].

Although this performance criterion was originally introduced for hard partitioning clustering, it is readily generalizable to soft partitioning. Indeed, if an object is assigned to clusters through a weighting scheme, these weights can replace the $\{0, 1\}$ entries in the co-membership matrix D and the n_{k_j} in Equation (2.10).

2.2.5 Curse of Dimensionality

Clustering analysis of microarray data may be applied in different settings depending on the objects to cluster, either the genes or the patients. The clustering of genes is not sensitive to the curse of dimensionality, since the dimensionality of the problem depends on the number of patients, which is usually smaller than the number of genes ($n \ll p$). However, the number of objects to cluster is large, and some of them may be irrelevant from a biological point of view (e.g. a cluster of non-expressed genes). In contrast, the clustering of patients suffers from the curse of dimensionality since the number of genes is larger than the number of patients ($p \gg n$, high feature-to-sample ratio). Therefore, clustering may be unstable and prone to overfitting [Everitt, 2002; Hastie et al., 2001] in this setting.

Whatever the objects to cluster (genes or patients), clustering analysis would benefit from a reduction in the number of genes. Indeed, the robustness of clustering methods is

usually improved by selecting only a subset of ”relevant” genes. As clustering methods are unsupervised, we should avoid the use of supervised feature selection methods.

In microarray analysis, feature selection methods based on unsupervised filtering are widely used. These unsupervised methods assess the quality of a gene through a criterion, such as its variance (genes with high variance may drive more relevant biological information) or its level of expression compared to a reference (large difference in gene expressions may be more biologically relevant) [Wessels et al., 2005; Saeys et al., 2007; Meyer, 2008]. This makes it possible to select a small subset of genes with the hope to remove noise and to improve the robustness of clustering methods.

2.2.6 Pitfalls and Dangers

Although clustering methods are unavoidable for the visualization and discovery of natural groupings in microarray data, they should be used with caution. Indeed, several issues have emerged since the first publications in the field. First, a clustering method always finds a structure in the data, depending on the choice of the method and the corresponding parameters (such as the number of clusters). Because there is no ”truth” or supervision (see Figure 2.7), it is hard to assess the quality of a clustering and to compare quantitively different methods (see Section 2.2.4). Second, due to the high feature-to-sample ratio of microarray data, clustering methods may be highly unstable, i.e. the identification of clusters may strongly depend on the data sample or the gene expressions under consideration. To reduce the risk of overfitting, most analysts attempt to reduce the feature-to-sample ratio by filtering the microarray data in order to remove noisy and uninformative genes (see Section 2.2.5). This filtering step, although beneficial, makes the clustering methods more complex to apply.

Additionally, analysts sometimes use clustering methods to perform classification, diverting the original purpose of such methods. In this case, the samples are labeled with respect to the clusters found in the dataset. The use of an unsupervised learning algorithm to perform a supervised task raises important issues, such as the optimization of performance: if a clustering method is used to classify a set of patients, it is not possible to optimize any performance criteria (e.g. sensitivity/specificity [Webb, 2003]) due to the fact that the clustering method does not use the supervised data (class of the patients) to build a model.

2.2.7 Concluding Remarks

At the time work related to this thesis was begun, clustering methods, especially hierarchical clustering, were widely used to perform class discovery and class prediction. The intrinsic characteristics of microarray data made difficult the validation of most of these results. We will show in Sections 4.2 and 5.3 how the integration of *a priori* biological knowledge for feature transformation improves the identification of breast cancer molecular subtypes through clustering analysis.

2.3 Survival Analysis

Survival Analysis is a class of statistical methods for studying the occurrence and timing of events [Allison, 1995]. These methods are most often applied to the study of deaths but can also handle different kinds of events, including the onset of disease and equipment

failure for instance. An event can be defined as a qualitative change[4] 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 time at which the event occurs is known. Some 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, with the time at which the events of interest occur being recorded. However, this observation plan is difficult to set up in practice, in that the investigator has to wait the end of follow-up before getting the final survival data. The alternative observation plan is *retrospective*, i.e. the survival data are retrieved from patients' medical histories. Yet these data present some potential limitations:

- The data are prone to errors; some events may be forgotten, especially when the duration of follow-up is long.
- The sample of patients may be a biased sample of the initial population of interest.

Survival data have a common feature, namely *censoring*, that is difficult to handle with conventional statistical methods. Consider the following example, which illustrates the problem of censoring. A sample of breast cancer patients were followed during 10 years after diagnosis. The event of interest was the appearance of a distant metastasis (a tumor initiated from the primary breast tumor cells and that is located in another organ). The aim was to determine how the occurrence and timing of distant metastasis appearance depended on several variables.

If we narrow our focus on a dichotomous dependent variable (free of distant metastasis or not), conventional methods that could analyze such data are, for example, logistic regression, linear discriminant analysis or support vector machines (see [Duda et al., 2001] for a review of such classification methods). But this sort of analysis ignores information on the timing of event. It is natural to suppose that patients who have a distant metastasis after 2 years have a more aggressive cancer than those who have distant metastasis after 9 years. 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 diagnosis and appearance of distant metastasis the dependent variable, and then estimate it by a conventional linear regression [McCullagh and Nelder, 1989]. But it remains a problem with patients who are free of distant metastasis during the 10 year period of follow-up. Such cases are referred to as *censored*. Two 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, and can result in large biases. Alternatively, the time of event could be set at 10 years for all those who are free of distant metastasis. This is clearly an underestimate, however, and some of those patients may never have a distant metastasis. Again, large biases may occur.

The methods of survival analysis allow for censoring by combining the information with the censored and the uncensored cases [Allison, 1995].

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

2.3.1 Censored 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 2.13. Suppose that this figure reports some of the data from a study in which all patients are diagnosed with a breast cancer at time $t = 0$ and are followed for 10 years thereafter. The horizontal axis represents time. Each of the horizontal lines labeled A through E represents a single patient. The symbol "**o**" indicates that a distant metastasis appeared at that point in time. The vertical line at 10 is the point at which the follow-up of the patients is stopped. Any distant metastases appearing at time 10 or earlier are observed and, hence, those occurrence times are uncensored. Any appearance occurring after 10 years are not observed, and those occurrence times are censored at time 10.

Patients A, C and D have uncensored occurrence times, while person B and E have right-censored occurence 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 10 years. It is just that their time of distant metastasis appearance did not exceed their censoring time. Moreover, the censoring time is fixed and is under the control of the investigator.

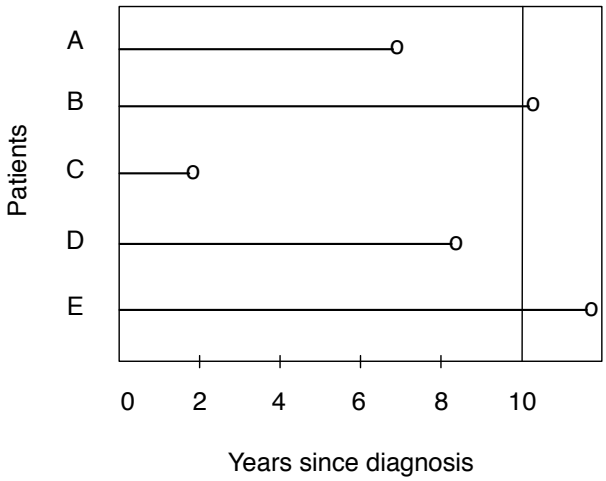


Figure 2.13: Singly right-censored data.

Random censoring occurs when observations are terminated for reasons that are not under the control of the investigator. This situation can be illustrated in our example. Patients who are still free of distant metastasis after 10 years are censored by a mechanism identical to that applied to the singly right-censored data. But some patients may move away, and it may be impossible to contact them. Some patients may die from another cause. Still other patients may refuse to participate after, say, 5 years. These kinds of censoring are depicted in Figure 2.14, where the symbol "**+**" for the patients A and C indicates that observation is censored at that point in time.

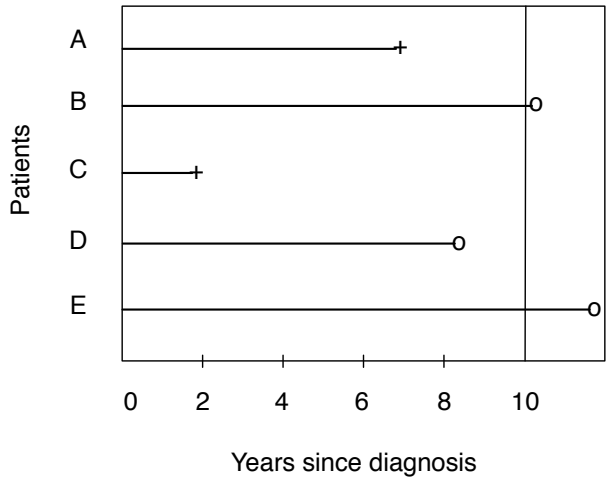


Figure 2.14: Randomly censored data.

Random censoring can also be produced when there is a single termination time, but entry times vary randomly across individuals. Consider again our example in which breast cancer patients are followed from diagnosis until the appearance of distant metastasis. A more likely scenario is one in which patients are diagnosed with breast cancer at various points in time, but the study has to be terminated on a single date. All patients still free of distant metastasis on that date are considered censored, but their survival times from diagnosis 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 t_i should not be affected if the individual is censored at t_i. That is, an individual who is censored at t should be representative of all those subjects with the same values of the explanatory variables who survive to t. (page 5)

The best way to understand this condition is to think about possible violations. In our example, it is plausible that those patients who refuse to continue participating in the study are more likely to be unsatisfied with their treatment because of cancer propagation and, hence, are at greater risk of distant metastasis. The censoring is informative assuming that measured explanatory variables do not fully account for the association between drop-out and cancer propagation. 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 the analysis of right-censored data with random (noninformative) censoring.

2.3.2 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: (i) the cumulative distribution function; (ii) the probability density function; and (iii) the hazard function.

2.3.2.1 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. $F(t) = \Pr\{\mathbf{t} \leq t\}$. In survival analysis, it is more common to work with the *survivor function*, defined as $S(t) = \Pr\{\mathbf{t} > t\} = 1 - F(t)$. If the event of interest is the appearance of a distant metastasis, the survivor function gives the probability of being free of metastasis until t. Because **t** cannot be negative, $S(0) = 1$.

2.3.2.2 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

$$f(t) = \frac{dF(t)}{dt} = -\frac{dS(t)}{dt} \tag{2.11}$$

Equivalently, we can write

$$f(t) = \lim_{\Delta t \to 0} \frac{\Pr\{t \leq \mathbf{t} < t + \Delta t\}}{\Delta t} \tag{2.12}$$

2.3.2.3 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

$$h(t) = \lim_{\Delta t \to 0} \frac{\Pr\{t \leq \mathbf{t} < t + \Delta t \,|\, \mathbf{t} \geq t\}}{\Delta t} \tag{2.13}$$

The function $h(t)$ quantifies the instantaneous risk that an event will occur in the small interval between t and $t+\Delta t$. The probability in the numerator of (2.13) 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 (2.13) is similar to an alternative definition of the PDF in Equation (2.12). The only difference is that the probability in the numerator is an unconditional probability, whereas the probability in (2.13) is conditional on $\mathbf{t} \geq t$. 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 Equation (2.11). Another simple formula expresses the hazard function in terms of the PDF and the survivor function:

$$h(t) = \frac{f(t)}{S(t)} \tag{2.14}$$

Together, (2.14) and (2.11) imply that

$$h(t) = -\frac{d}{dt} \log S(t) \tag{2.15}$$

By integrating both sides of (2.15), we obtain an expression of the survivor function in terms of the hazard function:

$$S(t) = \exp\left\{-\int_0^t h(u)du\right\} \tag{2.16}$$

Together with (2.14), this formula leads to

$$f(t) = h(t) \exp\left\{-\int_0^t h(u)du\right\} \tag{2.17}$$

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 having a distant metastasis at some particular point in time is 0.15, with time measured in years. This means that if the hazard stays at that value during a period of one year, one expects that a patient will have an distant metastasis 0.15 times during that year.

2.3.2.4 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 (2.16) and carrying out the integration implies that the survival function is $S(t) = e^{-\lambda t}$. From (2.11), 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:

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

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 (Figure 2.15). Alternatively we can assume that

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

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 (Figure 2.16).

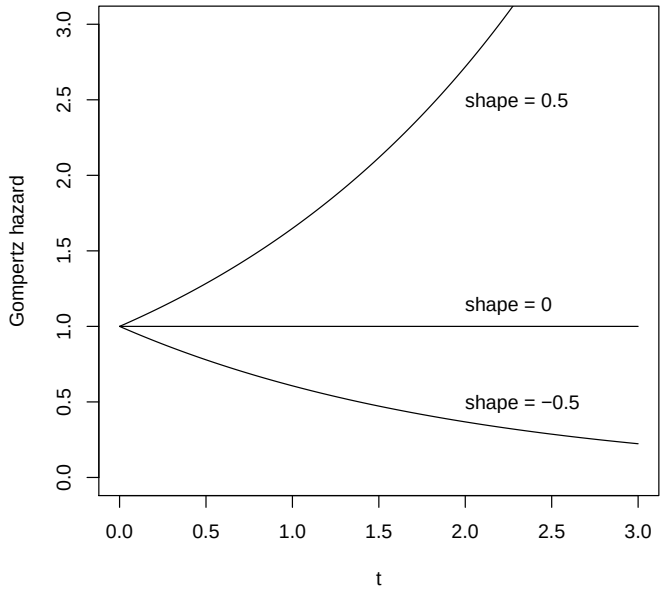


Figure 2.15: Typical hazard functions ($h(t) = \lambda e^{\alpha t}$ with $\lambda = 1$, α being the shape parameter) for the Gompertz distribution.

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 extend each of these models to allow for the linear influence of covariates. For instance, a covariate for the situation reported by the Figure 2.13 could be the age of the patient or tumor size at the time of diagnosis. Thus, if we have covariates $x_1, x_2, \ldots, x_p$, we can write

$$\textit{Exponential}: \quad \log h(t) = \mu + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \tag{2.18}$$

$$\textit{Gompertz}: \quad \log h(t) = \mu + \alpha t + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \tag{2.19}$$

$$\textit{Weibull}: \quad \log h(t) = \mu + \alpha \log t + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \tag{2.20}$$

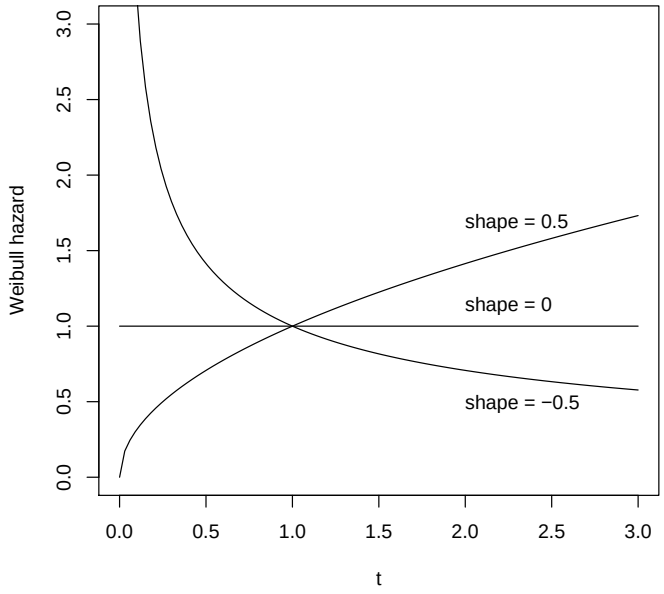


Figure 2.16: Typical hazard functions ($h(t) = \lambda t^{\alpha}$ with $\lambda = 1$, α being the shape parameter) for the Weibull distribution.

2.3.3 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 and for computing derived quantities from regression models (e.g. the median survival time or the five-year probability of survival).

There are two main methods to estimate survivor functions: the *Kaplan-Meier* (KM) and the *life-table* methods. The KM 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, we will use the KM estimator since the number of samples is usually small in survival analysis of microarray data.

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

When there is no censored data, the KM estimator is simple and intuitive. We have seen in Section 2.3.2 that the survivor function $S(t)$ is the probability that a time of event occurrence (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 time of event occurrence 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

$$\widehat{S}(t) = \prod_{j:t_j \leq t} \left[1 - \frac{d_j}{n_j}\right] \tag{2.21}$$

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. Consider the data in Table 2.2. The corresponding survival curve using the KM estimator is given in

Patient id	Survival time (years)	Event
1	1	1
2	2	0
3	4	1
4	5	0
5	5	1
6	7	1
7	8	0

Table 2.2: Example of survival times for breast cancer patients, the event being the appearance of a distant metastasis for instance.

Figure 2.17.

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

$$\begin{aligned} \hat{\sigma}^2_G\left\{\widehat{S}(t)\right\} &= \{\widehat{S}(t)\}^2 \sum_{j:t_j \leq t} \frac{d_j}{n_j(n_j - d_j)} \\ \hat{\text{se}}_G\left\{\widehat{S}(t)\right\} &= \widehat{S}(t)\sqrt{\sum_{j:t_j \leq t} \frac{d_j}{n_j(n_j - d_j)}} \end{aligned} \tag{2.22}$$

This is derived by estimating each term in the product expansion of $\widehat{S}(t)$ separately. Alternatively, 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 to the sample size, with mean $\widehat{S}(t)$ and variance estimated by the Greenwood formula [Meier, 1975]. Intervals of confidence around KM estimates can be computed using these results.

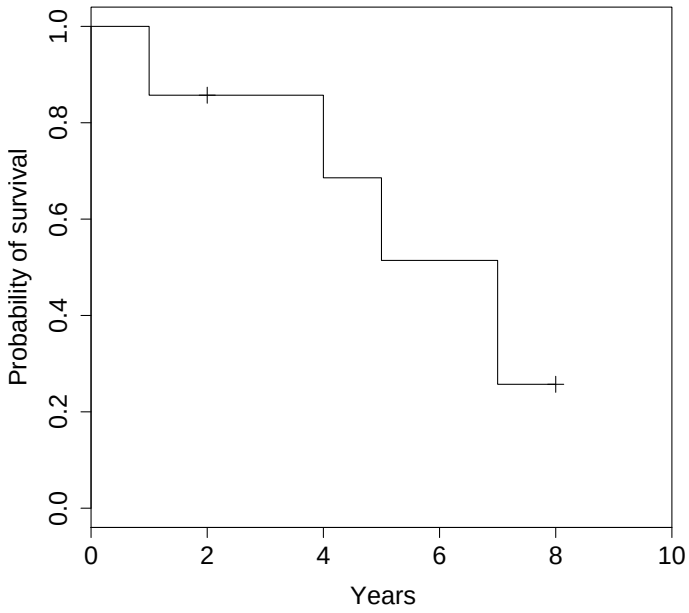


Figure 2.17: Survival curve estimated by the KM estimator from data in Table 2.2. The symbol "**+**" represents the censoring.

2.3.4 Estimating Regression Models

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

2.3.4.1 Parametric Regression Models

The class of parametric regression models is known as the *accelerated failure time* (AFT) class. In its 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 one individual different from another is the rate at which he or she ages. A good example is the conventional wisdom that a year for a dog is equivalent to seven years for a human.

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

$$\log \mathbf{t}_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i \tag{2.23}$$

where ϵ_i is a random disturbance term, and $\beta_0, \beta_1, ..., \beta_n$ 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. This is the case for one member of the AFT class, the *log-normal model*[5]. However, there are several alternatives that allow distributions of ϵ besides the normal distribution, but retain the assumptions of constant mean and variance, as well as independence across observations (see [Allison, 1995] for a description of the alternatives).

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

Recently, parametric regression models have been eclipsed by the semiparametric regression model, the renowned Cox regression model. This is why this thesis will focus on that method.

2.3.4.2 Semiparametric Regression Models

The semiparametric regression model refers to the method first proposed in 1972 by the British statistician Cox in his seminal 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 branches 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*.

Proportional hazards model Cox made two significant innovations. First, he proposed a model that is standardly referred to as the *proportional hazards model*. Second, he proposed a new estimation method that was later named *maximum partial likelihood*. The term *Cox regression* refers to the combination of the model and the estimation method.

Model The model is usually written as

$$h_i(t) = \lambda_0(t) \exp\left(\beta_1 x_{i1} + \cdots + \beta_p x_{ip}\right) \tag{2.24}$$

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 *p* 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 (2.24), we can rewrite the model as

$$\log h_i(t) = \alpha(t) + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} \tag{2.25}$$

[5]This model is called the log-normal model because if $\log \mathbf{t}$ has a normal distribution, then $\mathbf{t}$ has a log-normal distribution.

where $\alpha(t) = \log \lambda_0(t)$. If we further specify $\alpha(t) = \alpha$, we get the exponential model with covariates (see Equation (2.18)). If we specify $\alpha(t) = \alpha t$, we get the Gompertz model. Finally, if we specify $\alpha(t) = \alpha \log t$, we have the Weibull model. 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 whatsoever.

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 (2.24)

$$\frac{h_i(t)}{h_j(t)} = \exp\left\{\beta_1(x_{i1} - x_{j1}) + \cdots + \beta_p(x_{ip} - x_{jp})\right\} \tag{2.26}$$

so 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 hazard functions for any two individuals, the proportional hazards property implies that the functions should be strictly parallel as depicted in Figure 2.18.

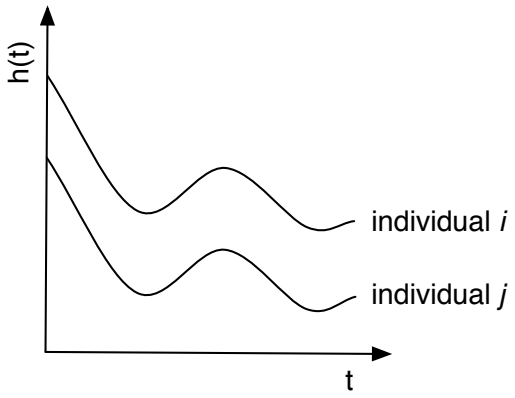


Figure 2.18: Parallel hazard functions from the proportional hazard model.

Estimation Fitting the proportional hazards model given in (2.24) to an observed set of survival data entails estimating the unknown coefficients, $\beta_1, \beta_2, \ldots, \beta_p$, of the covariates $x_1, x_2, \ldots, x_p$, 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 [Collett, 2003]. This is an important result, since it means that in order to make inferences about the effect of p covariates, $x_1, x_2, \ldots, x_p$, 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 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 still have 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 of three parts: t_i, δ_i and x_i, where t_i is the time of the event occurrence 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 $x_i = [x_{i1}, x_{i2}, \ldots, x_{ip}]$ is a vector of p 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

$$PL = \prod_{i=1}^{n} L_i \tag{2.27}$$

where L_i is the likelihood for the ith event. Next we need to know how the L_i are constructed. This is best explained by way of an example. Consider the data in Table 2.2 where we add a column 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).

Patient id	Survival time (years)	Event	x
1	1	1	1
2	2	0	1
3	4	1	1
4	5	0	0
5	5	1	1
6	7	1	0
7	8	0	0

Table 2.3: Example of survival times for breast cancer patients with the covariate x.

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

$$L_1 = \frac{h_1(1)}{h_1(1) + h_2(1) + \cdots + h_7(1)} \tag{2.28}$$

The second event occurred to patient 3 in year 4. Patient 1 is no longer at risk of event because he or she already had an event before year 4. Patient 2 was not longer at risk because he or she was censored at year 2. So L_3 has the same form as L_1, but the hazards for patient 1 and 2 are removed from the denominator:

$$L_3 = \frac{h_3(4)}{h_3(4) + \cdots + h_7(4)} \tag{2.29}$$

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

The set of all individuals who are at risk at a given point in time is often referred to as the *risk set*. At year 4, the risk set consists of patients 3 to 7.

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

$$L_i = \left[\frac{e^{\beta X_i}}{\sum_{j=1}^{n} y_{ij} e^{\beta X_j}} \right]^{\delta_i} \tag{2.30}$$

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 at t_i is excluded because $\delta_i = 0$ for those cases[7].

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

$$PL = \prod_{i=1}^{n} \left[\frac{e^{\beta X_i}}{\sum_{j=1}^{n} y_{ij} e^{\beta X_j}} \right]^{\delta_i} \tag{2.31}$$

Once the partial likelihood is constructed, it can be maximized with respect to β just like an ordinary likelihood function. It is convenient to maximize the logarithm of the likelihood, which is

$$\log PL = \sum_{i=1}^{n} \delta_i \left[\beta X_i - \log \left(\sum_{j=1}^{n} y_{ij} e^{\beta X_j} \right) \right] \tag{2.32}$$

Most partial likelihood programs use some version of the Newton-Raphson algorithm [Collett, 2003] to maximize this function with respect to β.

The formula 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, assuming 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.

Stratification An extension of Cox's model allows for multiple *strata*. The strata divide the individuals into disjoint groups[8], each of which has a distinct baseline hazard function but common values for the coefficients β. Assume that individuals $i = 1, \ldots, n_1$ are in stratum 1, individuals $n_1, \ldots, n_1 + n_2$ are in stratum 2, and so on until stratum s. The hazard for an individual i, who belongs to stratum k is

$$h_i(t) = \lambda_k(t) \exp(\beta x_i)$$

Analysis of multicenter clinical studies frequently uses stratification. Because of varying patient populations and referral patterns, the different centers in a study are likely to have different baseline hazards, ones that do not have the simple parallel relationship shown in Figure 2.18.

[7] The censored information are taken into account to estimate the likelihood L until their censoring time is reached. This ensures to make full use of the survival data to estimate the likelihood.

[8] Because they are used to divide the individuals into a disjoint set of groups, stratification variables are effectively treated as categorical.

Computationally, the overall log partial likelihood in (2.32) becomes a sum

$$\log PL = \sum_{k=1}^{s} \log PL_k$$

where $\log PL_k$ is precisely Equation (2.32), but summed over only the individuals in stratum k.

The main advantage of stratification is that it gives the most general adjustment for confounding variables. The disadvantage is that no direct estimation of the importance of the strata effect is produced.

2.3.5 Performance Assessment

The output of a survival model (prediction) depends on the class of the model: for parametric models, the time of event occurrence is estimated; for the semi-parametric Cox's model, the hazard is estimated. Since Cox's model is widely used, we will focus on the performance of this model.

The coefficients β in Cox's model are estimated without the baseline hazard function $\lambda_0(t)$ (Section 2.3.4.2). Since all the patients have the same baseline hazard function, the order of the patients with respect to their hazard function $h(t)$ depends only on the linear combination βx, also called *risk score* in the literature. This risk score can be transformed into hazard or survival probability using Equation (2.16) by estimating separately the baseline hazard function $\lambda_0(t)$ [Therneau and Grambsch, 2000].

Clinicians often use the predicted risk scores, hazards or survival probabilities to derive *risk groups* through the application of one or several cutoffs. Although the discretization of individual risk scores into a finite (and often small) set of risk groups may introduce bias [Gerds and Schumacher, 2001; Royston et al., 2006], this approach is very intuitive and conforms to the daily decision making process of doctors, e.g. the attribution of either low or high risk to patients.

We will present in the next sections several performance criteria to assess the accuracy of a survival model through their risk score and risk group predictions. In the following the quantity r_i and g_i will denote the risk score and the risk group for patient i, respectively. r is a real value and g is either 0 or 1 for a low- or high-risk patient, respectively[9].

Properties of a "good" performance criterion Royston and Sauerbrei defined a set of properties that a "good" performance criterion should possess [Royston and Sauerbrei, 2004]. The properties are as follows:

Interpretability: The performance criterion should have a simple and intuitively appealing meaning.

Generality: The performance criterion should be applicable to risk score and risk group predictions.

Directedness: When the risk ordering changes, the performance criterion should change in the appropriate direction. For example, if the risk ordering were reversed then the performance criterion should change sign.

[9] All the formula presented hereafter are easily generalizable to more than two risk groups.

Unbiasedness: The observed value of the performance criterion should be an unbiased estimate of the true value. In particular, the expectation of the performance criterion should be close to a known value when the proposed risk ordering bears on average no relationship with the true risk ordering. This may occur when the model is useless, or less commonly, when the dadasets holds no information relevant for the prediction task.

Responsiveness: If the risk gradient were reduced, for example by omission of a covariate highly relevant for the prediction task, then the performance criterion should move substantially towards a known value.

Robustness: The performance criterion should not be unduly influenced by a small number of extreme risk predictions (outliers).

Precision: Confidence intervals for the performance criterion should be computed straightforwardly.

Independence of censoring: The performance criterion should be independent to the degree of censoring (within the time frame of interest) in the data.

It is worth noting that all the performance criteria presented below are not independent of the degree of censoring in survival data. However, the resulting bias is usually small (see publications specific to each performance criterion).

2.3.5.1 Risk Score Prediction

In this section, we will present an inventory of performance criteria used for risk score prediction. It should be observed that, due to the presence of censoring, there is no equivalent in survival analysis to the concept of mean squared error, widely used in linear regression [van Houwelingen et al., 2006]. This is the reason why different performance criteria were developed to specifically deal with censoring.

Cross-validated partial likelihood This performance criterion was originally introduced in penalized Cox regression to optimize the penalization term [Verweij and van Houwelingen, 1993]. However, it can be used to assess the performance of a risk prediction through the computation of the partial log likelihood of a Cox model in a cross-validation framework.

Considering a Cox model with the risk score as input variable, $h(t) = \lambda_0(t)\exp(\beta r)$, the cross-validated partial likelihood, denoted by CVPL, in a leave-one-out cross-validation framework is defined as

$$\begin{aligned} CVPL &= -\sum_{i=1}^{n} l_i\left(\hat{\beta}^{(-i)}\right) \\ &= -\sum_{i=1}^{n}\left[l\left(\hat{\beta}^{(-i)}\right) - l^{(-i)}\left(\hat{\beta}^{(-i)}\right)\right] \end{aligned} \tag{2.33}$$

where $\hat{\beta}^{(-i)}$ is the partial likelihood estimate of the coefficient from the data without the i^{th} individual. The terms $l(\beta)$ and $l^{(-i)}(\beta)$ are the log partial likelihoods with all the individuals

and without the i^{th} individual, respectively. The term $l_i(\beta)$ is the contribution of individual i to the log partial likelihood at β.

The risk prediction should maximize the sum of the contributions of each individual to the log partial likelihood, and thus minimize the CVPL. The interpretation of the CVPL values is dependent on the dataset under study. However we can normalize these values with respect to the CVPL of the null model, i.e. $\beta = 0$. The normalized CVPL, denoted by $CVPL_{norm}$, is defined as

$$CVPL_{norm} = \frac{\sum_{i=1}^{n} \left[l\left(\hat{\beta}^{(-i)}\right) - l^{(-i)}\left(\hat{\beta}^{(-i)}\right) \right]}{\sum_{i=1}^{n} \left[l(0) - l^{(-i)}(0) \right]} \tag{2.34}$$

where $l(0)$ and $l^{(-i)}(0)$ are the log partial likelihoods of the null model with all the individuals and without the i^{th} individual, respectively.

It is worth noting that the CVPL is not restricted to leave-one-out cross-validation and can easily be extended to other cross-validation frameworks.

Standard error To the best of our knowledge, there is no reference in the literature to the standard error of the CVPL.

Properties $CVPL_{norm}$ possesses the properties of generality and responsiveness.

Time-dependent ROC curve The receiver operating characteristic (ROC) curve is a standard technique for assessing the performance of a continuous variable for binary classification [Sweets, 1988]. A ROC curve is a plot of sensitivity versus 1 − specificity for all the possible cutoff values of the continuous variable, denoted by c. In survival analysis, the continuous variable is the risk score, and the binary class to predict is the event occurrence, denoted by $d(t)$. As the event occurrence is time-dependent, time-dependent ROC curves are more appropriate than conventional ones. In [Heagerty et al., 2000], the authors proposed to summarize the discrimination potential of a risk score r, estimated at the diagnosis time $t = 0$, by calculating ROC curves for cumulative event occurrence by time t. Once we define the sensitivity SE and the specificity SP as follows

$$SE(c, t, r) = \Pr\{r > c \mid d(t) = 1\} \tag{2.35}$$

$$SP(c, t, r) = \Pr\{r \leq c \mid d(t) = 0\} \tag{2.36}$$

the ROC curve $ROC(t)$ at time t is the plot of $SE(c, t, r)$ versus $1 - SP(c, t, r)$ where the cutoff point c is the parameter. In order to estimate the conditional probabilities in (2.35) and (2.36), accounting for possible censoring, the nearest neighbor estimator for the bivariate distribution function proposed by [Akritas, 1994] is used in preference to the KM estimator. Indeed the KM estimator does not guarantee that sensitivity and specificity are monotone (see [Heagerty et al., 2000] for an example).

From the time-dependent ROC curve $ROC(t)$ we can summarize the performance of a risk score by deriving the area under the curve quantity, denoted by $AUC(t)$. Since AUC depends on time t, we define the *integrated area under the curve* (IAUC) as the area under $AUC(t_i), \forall i$ such that $\delta_i = 1$.

Both $AUC(t)$ and $IAUC$ lie in $[0, 1]$, the performance of the risk score produced by a random model being equal to 0.5. The performance increases as the departure from 0.5 increases. See Figure 2.19 for ROC curves example.

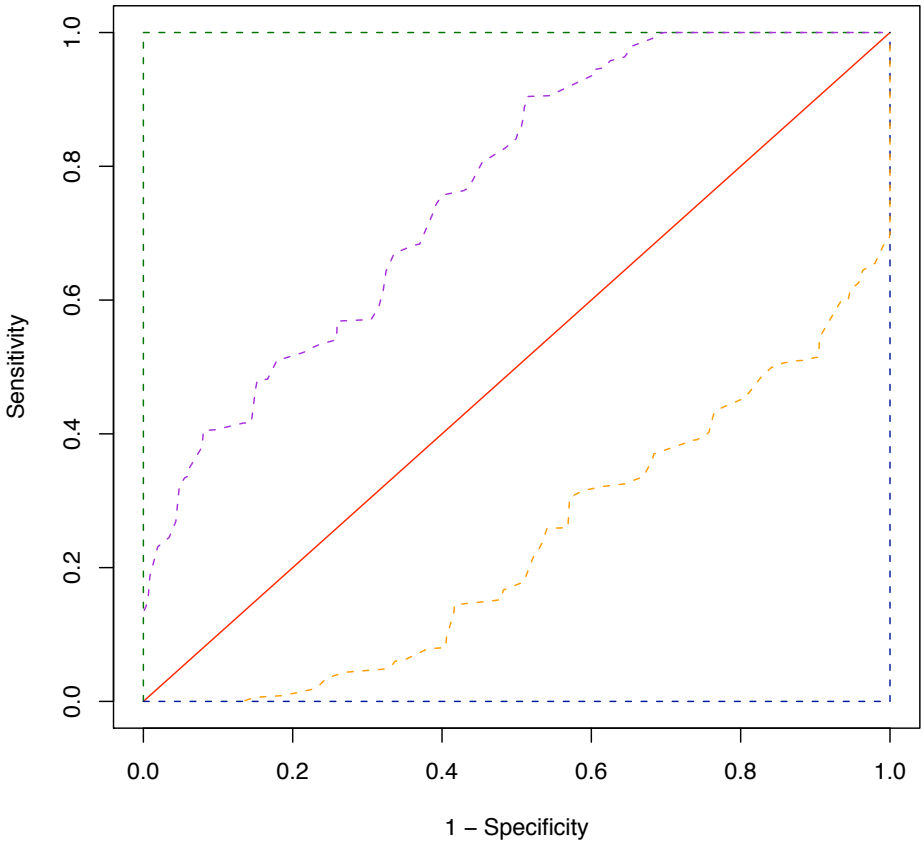


Figure 2.19: Example of ROC curves. The red diagonal line represents the performance of the risk score of a random model. The green and violet curves represent the performance of perfect and non perfect risk scores, respectively, such that large risk scores stand for high-risk patients. The blue and orange curves represent the performance of perfect and non perfect risk scores, respectively, such that large risk scores stand for low-risk patients.

Standard error To the best of our knowledge, there is no reference in the literature to the standard error of the $AUC(t)$ or the $IAUC$.

Properties $AUC(t)$ and $IAUC$ possess the properties of interpretability, directedness, unbiasedness, responsiveness and robustness.

Concordance index The concordance index (C-index) computes the probability that, for a pair of randomly chosen comparable patients (see below), the patient with the higher risk prediction will experience an event before the lower risk patient (or inversely). The C-index takes the form

$$C\text{-index} = \frac{\sum_{i,j\in\Omega} 1\{r_i > r_j\}}{|\Omega|} \tag{2.37}$$

where r_i and r_j stand for the risk predictions of the i^{th} and the j^{th} patient, respectively, and Ω is the set of all the pairs of patients $\{i, j\}$ for whom there is no tie in risk predictions ($r_i \neq r_j$) and who meet one of the following conditions: (i) both patients i and j experienced an event and time $t_i < t_j$ or (ii) only patient i experienced an event and $t_i < c_j$.

Note that the C-index is a generalization of the $AUC(t)$ (with similar interpretation), though it is unable to represent the evolution of performance with respect to time [Harrell et al., 1996].

Standard error Standard error, confidence intervals and p-values for the C-index are computed by assuming asymptotic normality [Pencina and D'Agostino, 2004].

Properties C-index possesses all the properties of a "good" performance criterion.

Brier Score The Brier score, denoted by BSC, is defined as the squared difference between an event occurrence and its predicted probabilities at time t. Probabilities of event, denoted by Q, can be derived from Cox's proportional hazards model fitted with the risk score r or risk group g predictions[10]. Intuitively, if a patient experiences no event at time t, the predicted probability of event occurrence should be close to zero. Symmetrically if the patient experiences an event, the probability should be close to one. The BSC formalizes this intuition by computing the time dependent quantity

$$BSC(t) = \sum_{i=1}^{n} (d_i(t) - q_i(t))^2 W \tag{2.38}$$

where the weights W are used to remove a large sample censoring bias [Graf et al., 1999; Gerds and Schumacher, 2006].

A summary of the predictability error over time is provided by the integrated Brier score, denoted by IBSC. Note that the lower the BSC, the better the predictability of patients' risks at time t. Similarly, the lower the IBSC, the better the average predictability of patients' risks.

[10] As the computation of probabilities of event occurrence requires the estimation of the baseline hazard function $\lambda_0(t)$, the Brier score is rarely used to assess the performance of a survival model.

For judging the (I)BSC, we usually rely on the score of a benchmark risk prediction model that is obtained with the overall Kaplan-Meier estimator [Kaplan and Meier, 1958] for the survival function. This simple risk prediction model corresponds to a model that assigns the same risk prediction to all patients. It ignores the information contained in explanatory variables completely and thus provides a suitable benchmark value similar to the one obtained with the null model in linear regression.

Standard error To the best of our knowledge, there is no reference in the literature to the standard error of the $BSC(t)$.

Properties $BSC(t)$ and *IBSC* possess the properties of generality and responsiveness.

D index The D index, denoted by D-index, is a measure of separation between the hazard function of each patient [Royston and Sauerbrei, 2004]. It is based on an estimation of the underlying spread of the log hazard ratios compared with the baseline hazard function in Cox's model (see Section 2.3.4.2)

From (2.25), Cox's model can be written as

$$\log h_i(t) = \log \lambda_0(t) + s_i \tag{2.39}$$

where $s_i = \beta_i r_i$ with r_i, the risk score for patient i.

Consider the distribution of the s values. Defining order statistics $s_{(1)} \leq \cdots \leq s_{(n)}$, we can generally write

$$s_{(i)} = \mu + \sigma u_i + \epsilon_i$$

where u_i is the i^{th} expected standard Normal order statistic (rankit) in a sample size n [Blom, 1958]. Ordering the data on the s_i and substituting for $s_{(i)}$ in (2.39), we have

$$\log h_i(t) = \log \lambda_0(t) + \mu + \sigma u_i + \epsilon_i$$

So far we have assumed no specific distribution for the s_i . Let us now suppose that $s_i \sim \mathcal{N}(\mu, \sigma^2)$. The parameter σ is the standard deviation of the s_i values and is a natural measure of separation. By definition, the regression of the $s_{(i)}$ on the u_i is linear with $E(s_{(i)}) = \mu + \sigma u_i$ and $E(\epsilon_i) = 0$.

To a first approximation, let set $\epsilon_i = 0$. Then

$$\log h_i(t) \approx \log \lambda_0(t) + \mu + \sigma u_i \tag{2.40}$$

Under the normality assumption, Cox's model (2.40) is approximately linear in each u_i. On fitting it to the data, the constant μ is absorbed into the baseline hazard function $\lambda_0(t)$ and the regression coefficient $\hat{\sigma}$ will estimate σ. The D index is defined as

$$D\text{-index} = \exp(\kappa\hat{\sigma})$$

where $\kappa = \sqrt{8/\pi}$.

Let $z_i = \kappa^{-1} u_i$ where the z_i corresponding to the tied values in $s_{(i)}$ are averaged[11]. The regression of Cox's model (2.40) on the z_i instead of the u_i, estimates D directly.

[11] This is particularly important for risk group predictions where each group is represented by tied values.

Scaling the u_i by κ lends a direct interpretation to D, as follows. Asymptotically, the mean of the positive z_i is 0.5 and the mean of the negative z_i is -0.5, deducible from the fact that the mean of a standard half-normal distribution is $\sqrt{\frac{2}{\pi}}$, i.e. $\frac{1}{2}\kappa$. Now suppose the $s_{(i)}$ are dichotomized by applying a cutoff at the median, or equivalently to the $z_i = 0$. Cox's regression on the group-averaged z_i (with values = ± 0.5) provides the same regression coefficient as Cox's regression on a binary dummy variable distinguishing the groups (see *hazard ratio* in the next section). Therefore, D is an estimate of the log hazard ratio comparing two equal-sized risk groups. This is a natural measure of separation between two independent survivor functions under the proportional hazards assumption.

The main advantage of the D-index compared to the hazard ratio is that the transformation of the risk scores in ranks allows for comparing performance of different datasets without calibration since the risk scores have the same scales. Moreover, the D-index is robust to outliers.

Standard error Once D is estimated through Cox's model, the corresponding confidence interval can be obtained from the standard error of $\hat{\beta} = \kappa\hat{\sigma}$ (see Section 2.3.4.2). So, a $(100 - \alpha)$% confidence interval for the true D 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].

Properties D index possesses all the properties of a "good" performance criterion.

2.3.5.2 Risk Group Prediction

In this section, we will present an inventory of performance criteria used for risk group prediction. It should be noted that, although hypothesis testing does not allow for quantifying the performance of a risk group prediction, it brings some insights into the significance of this prediction compared to the performance of a null model (usually the random case).

Hypothesis testing Since the number of predicted risk groups is usually small, the corresponding survivor functions (or survival curves) can be estimated. Testing for differences in survivor functions is an important topic in survival analysis. For instance, if two groups of patients are defined by a metastasis marker (appearance of metastasis or not), the obvious question to ask is "Did the two groups exhibit different survival?". Since the survivor function gives a complete accounting of the survival experience of each group, a natural approach for 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 are several alternative statistics for testing this null hypothesis. We will present here the logrank test (also known as the Mantel-Haenzel test), since this test is the most widely used in this setting.

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 j^{th} 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

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

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

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

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

$$U_L = \sum_{j=1}^{r}(d_{1j} - e_{1j}) \tag{2.41}$$

Since the event times are independent of one another, the variance of (2.41) 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

$$\text{var}(D_{1j}) = \frac{n_{1j}(n_j - n1j)d_j(n_j - d_j)}{n_j^2(n_j - 1)} \tag{2.42}$$

so that the variance of U_L is

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

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 with one degree of freedom, denoted χ^2_1, and so we have that

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

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

The logrank test readily generalizes to three or more groups, with the null hypothesis that all groups have the same survivor function. If the null hypothesis is true, the test statistic has a chi-square distribution with a degree of freedom equal to the number of groups minus 1.

Hazard ratio The hazard ratio can be defined as a summary of the difference between two survival curves, representing the reduction in the risk of event between two different groups. It is a form of relative risk. A proportional hazards regression model assumes that the relative risk of event between the two groups is constant at each interval of time.

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

$$h_i(t) = \lambda_0(t) \exp(\beta g_i) \tag{2.43}$$

where $g_i = 1$ if the i^{th} 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 in the first group. Moreover, the hazard function for any individual in the second group is $\psi\lambda_0(t)$ (proportional hazards). ψ is the relative hazard or *hazard ratio* (HR) 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 us to assess whether the risk of the high-risk group is higher than in the low-risk group.

Standard error Once the parameter β is estimated, giving $\hat{\beta}$, the corresponding estimate of the hazard ratio is $\hat{\psi} = \exp(\hat{\beta})$. The confidence interval of $\hat{\psi}$ can be obtained from the standard error of $\hat{\beta}$ (see Section 2.3.4.2). So 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].

Properties HR possesses the properties of generality, directedness, responsiveness and precision.

Other performance criteria The cross-validated partial likelihood, the concordance index, the Brier score and the D index are applicable for risk group prediction as well, r_i being replaced by g_i in the formula.

2.3.6 Curse of Dimensionality

As for clustering methods (see Section 2.2), the regression of survival models (e.g. Cox's regression) is sensitive to the curse of dimensionality. Since the number of genes is larger than the number of samples ($p \gg n$, high feature-to-sample ratio), the quality of the survival models would benefit from the use of a reduced subset of relevant genes. In addition to the filtering step described in Section 2.2.5, a supervised feature selection is usually performed. This allows the analyst to select a small subset of genes relevant for the prediction of patients' survival. However, due to the high dimensionality of microarray data, the high level of noise and the correlation between variables (features) due to gene co-expressions (see Section 2.1.2), the feature selection is a difficult task and is often referred to as a key step in microarray data analysis (Section 2.1.3.2).

2.3.7 Pitfalls and Dangers

The main pitfall of survival analysis is the lack of a gold standard for performance assessment of survival models. An inventory of performance criteria for risk score and risk group prediction was presented in Sections 2.3.5.1 and 2.3.5.2, respectively. Authors usually have their performance criterion of preference, making difficult the comparison of results from different publications.

The main danger in the survival analysis of microarray data is the curse of dimensionality as explained in the previous section. In addition to the computational cost to fit a survival model with many genes, overfitting [Everitt, 2002; Hastie et al., 2001] usually prevents the use of large scale multivariate models. Although we presented only linear survival models, several authors have recently introduced non-linear survival models with some application to microarray data analysis [Ripley et al., 2004; Molinaro et al., 2004; van Belle et al., 2007]. The aforementioned risk of overfitting also (even more) applies to such non-linear survival models.

2.3.8 Concluding Remarks

Although survival analysis has a long history in epidemiology, psychology and clinical studies, these methods are not widely used in microarray data analysis, especially for breast cancer prognosis. When this thesis was begun, the field lacked both a robust methodology for signature extraction and a large comparative study to uncover the key characteristics of successful risk prediction models.

Chapter 3

State-of-the-Art

Clinicians have long recognized that breast cancer is a heterogenous disease, breast tumors exhibiting different histo-pathological characteristics (ER status or histological grade, see Section 1.2.1). However, traditional histo-pathological characteristics (Section 1.2.1) are unable to capture the biologic heterogeneity of breast tumors, which makes it difficult to assess the prognosis of patients. Indeed, patients having similar tumors in terms of histo-pathological characteristics may exhibit dramatically different clinical outcome.

Clinical investigators have put great hope in gene expression profiling technologies and the new type of data generated as a consequence, and view them as a means to improve our understanding of breast cancer at the molecular level and to improve the traditional prognostic models. These goals may be achieved by studying the molecular heterogeneity of breast tumors and by identifying molecular markers exhibiting high prognostic value.

In the next sections, we will present the state-of-the-art of the following:

- The knowledge about the key biological processes involved in breast cancer tumorigenesis.
- The discovery of breast cancer molecular subtypes, which allow clinicians to better understand which genetic traits are shared by tumors.
- The identification of global prognostic gene signatures, i.e. signatures extracted from the whole dataset without considering the presence of different molecular subtypes.
- The identification of local prognostic gene signatures, i.e. signatures extracted from specific molecular subtypes present in the whole dataset.
- The performance assessment of the prognostic models built from the global and local gene signatures, and the comparison of the performance of traditional models versus competitive gene signatures.

3.1 Breast Cancer Biology

After a quarter century of rapid advances, cancer research has generated a rich and complex body of knowledge, revealing cancer to be a disease involving dynamic changes in the

genome. Several lines of evidence indicate that tumorigenesis in humans is a multistep process and that these steps reflect genetic alterations that drive the progressive transformation of normal human cells into highly malignant derivatives [Hanahan and Weinberg, 2000].

Hanahan and Weinberg suggest that the vast catalog of cancer cell genotypes is a manifestation of a small set of essential alterations in cell physiology that collectively dictate malignant growth [Hanahan and Weinberg, 2000]. In breast cancer, the key biological processes involved in tumorigenesis (Figure 3.1) are the following:

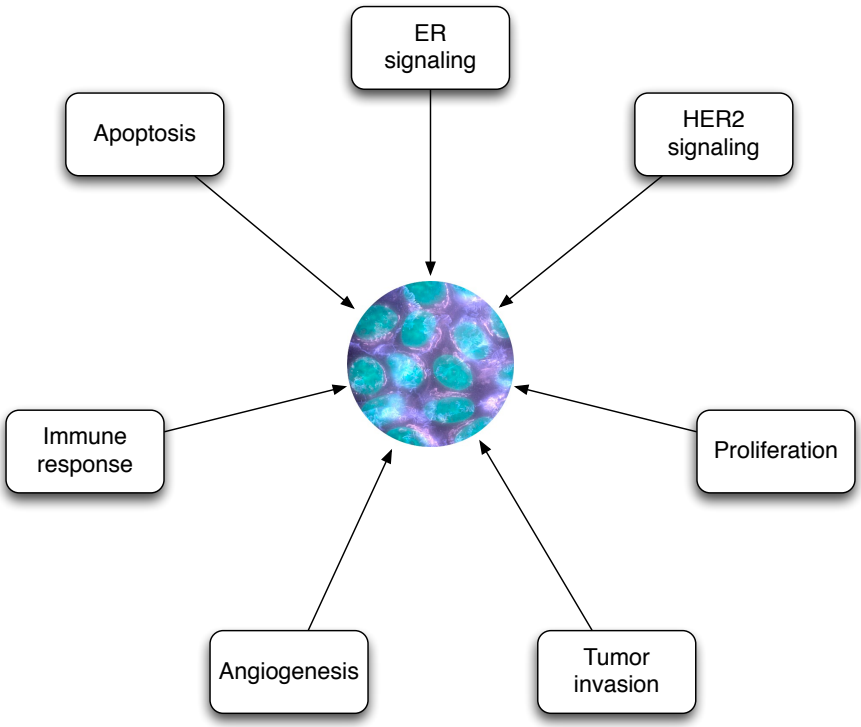


Figure 3.1: Key biological processes involved in breast tumorigenesis [Hanahan and Weinberg, 2000]. The arrows are drawn for presentation purpose and do not indicate the strength of the relation between the biological processes and tumorigenesis. Actually, the biological processes have different impact on tumor progression and are highly interconnected but these relations are barely known.

ER signaling: The estrogen receptor (ER) is a protein found inside the cells of the female reproductive tissue, some other types of tissue, and some cancer cells. The hormone estrogen will bind to the receptors inside cells and may cause the cells to grow. The majority of breast cancer cells are ER-positive and need estrogen to grow, and may stop growing when treated with hormones that block estrogen from binding.

HER2 signaling: HER2, also called c-erbB-2 or human epidermal growth factor receptor2, is a protein involved in normal cell growth. It is found in high levels on some breast cancer cells and its overexpression has been found to correlate with more aggressive forms of the disease.

Proliferation: Proliferation represents an increase in the number of cells as a result of cell growth and cell division. Uncontrolled cell proliferation is one of the major hallmarks of cancer and it has been widely investigated in breast cancer for its association with neoplastic growth, progression, and metastatic potential.

Tumor invasion: Tumor invasion occurs when the tumor spreads beyond the layer of tissue in which it developed and grows into surrounding, healthy tissues. It involves changes in the physical coupling of cells to their microenvironment and activation of extracellular proteases.

Angiogenesis: Tumor angiogenesis is the proliferation of a network of blood vessels that penetrates into cancerous growths, supplying nutrients and oxygen and removing waste products. Tumor angiogenesis actually starts with cancerous tumor cells releasing molecules that send signals to surrounding normal host tissue. This signaling activates certain genes in the host tissue that, in turn, makes proteins to encourage growth of new blood vessels.

Immune response: The immune response is the activity of the immune system against foreign substances (antigens). When normal cells turn into cancer cells, some of the antigens on their surface change. These cells, like many body cells, constantly shed bits of protein from their surface into the circulatory system. Often, tumor antigens are among the shed proteins. These shed antigens prompt action from immune defenders, including cytotoxic T cells, natural killer cells, and macrophages. According to the theory, patrolling cells of the immune system provide continuous bodywide surveillance, catching and eliminating cells that undergo malignant transformation. Tumors then develop when this immune surveillance system breaks down or is overwhelmed.

Apoptosis: Apoptosis is a type of cell death in which a series of molecular steps in a cell leads to its death. This is the body's normal way of getting rid of unneeded or abnormal cells. The process of apoptosis may be blocked in cancer cells. It is also called *programmed cell death*.

Each of these physiologic changes – novel capabilities acquired during tumor development – represents the successful breaching of an anticancer defense mechanism hardwired into cells and tissues.

3.2 Breast Cancer Molecular Subtypes

Since breast tumors are biologically heterogeneous and exhibit different clinical outcomes (Section 1.1.1), an accurate identification of molecular subtypes would make it possible to better understand breast cancer biology and to test the prognostic value of molecular markers with respect to these subtypes.

Bioinformatics studies dealing with the identification of breast cancer subtypes from gene expression data (unsupervised learning, see Section 2.1.3.3) are usually confronted with the following difficult choices:

- Which clustering method?
- Which subset of genes?

- How many clusters?
- How to use the clustering model to classify new cases?
- How to validate the robustness of the clustering model?

Initial studies [Perou et al., 2000; Sorlie et al., 2001, 2003] used the hierarchical clustering method (Section 2.2.1) to highlight the presence of natural groupings of breast tumors. We illustrate the method used in these initial publications in Figure 3.2. The authors selected a subset of several hundreds of highly variant genes (e.g. the "intrinsic gene list", first mentioned in [Perou et al., 2000]). After performing hierarchical clustering, the dendrogram was cut to identify the different subtypes, based on a subjective visualization assessment (Figure 3.2). As such, the hierarchical clustering model fitted onto the training set could not be used directly to identify the subtype of a tumor of a new breast cancer patient. Indeed, any new case should be added to the training set and the hierarchical clustering model should be fitted again, leading to a potentially different dendrogram. To circumvent this difficulty, the authors developed a method based on *nearest centroid* [Dudoit et al., 2002], called SSP (Single Sample Predictor, see [Sorlie et al., 2003]): first, a mean gene expression profile (called *centroid*) was created for each subtype (see Figure 3.2 for an example); second, the gene expression profile of the new case was compared to each centroid and assigned by the SSP to the nearest subtype centroid as determined by Spearman correlation (Figure 3.2).

We illustrate hereafter the use of this procedure in [Sorlie et al., 2001]. We can summarize the new insights into breast cancer biology introduced by this study with the following findings:

- ER and HER2 signaling pathways (see previous section for details) were shown to have the strongest association with the gene expression profile of breast tumors. Sorlie et al. stated that, although the intrinsic gene list includes several unknown genes, the clustering was mainly driven by the genes related to ER (ESR1) and HER2 (ERBB2) signaling pathways (Figure 3.3).
- Breast tumors can be grouped into at least four subtypes; these are the basal-like (mainly ER- and HER2-), the HER2+ and two to three luminal (mainly ER+ and HER2-, characterized by different expression levels of proliferation genes) as sketched in Figure 3.4.
- Each subtype exhibits a distinct clinical outcome, i.e. a different natural history or response to various treatments (Figure 3.5).

Although these initial results were promising, several issues remained open [Pusztai et al., 2006]:

- The use of hierarchical clustering did not allow for the easy classification of a new patient and did not provide an accurate estimate of the classification uncertainty. Sorlie et al. addressed the former issue by developing the single sample predictor (SSP) but the later remained open.
- The use of a large number of genes – the intrinsic gene list contains more than 500 genes – might lead to a clustering model prone to overfitting (low robustness due to high feature-to-sample ratio).

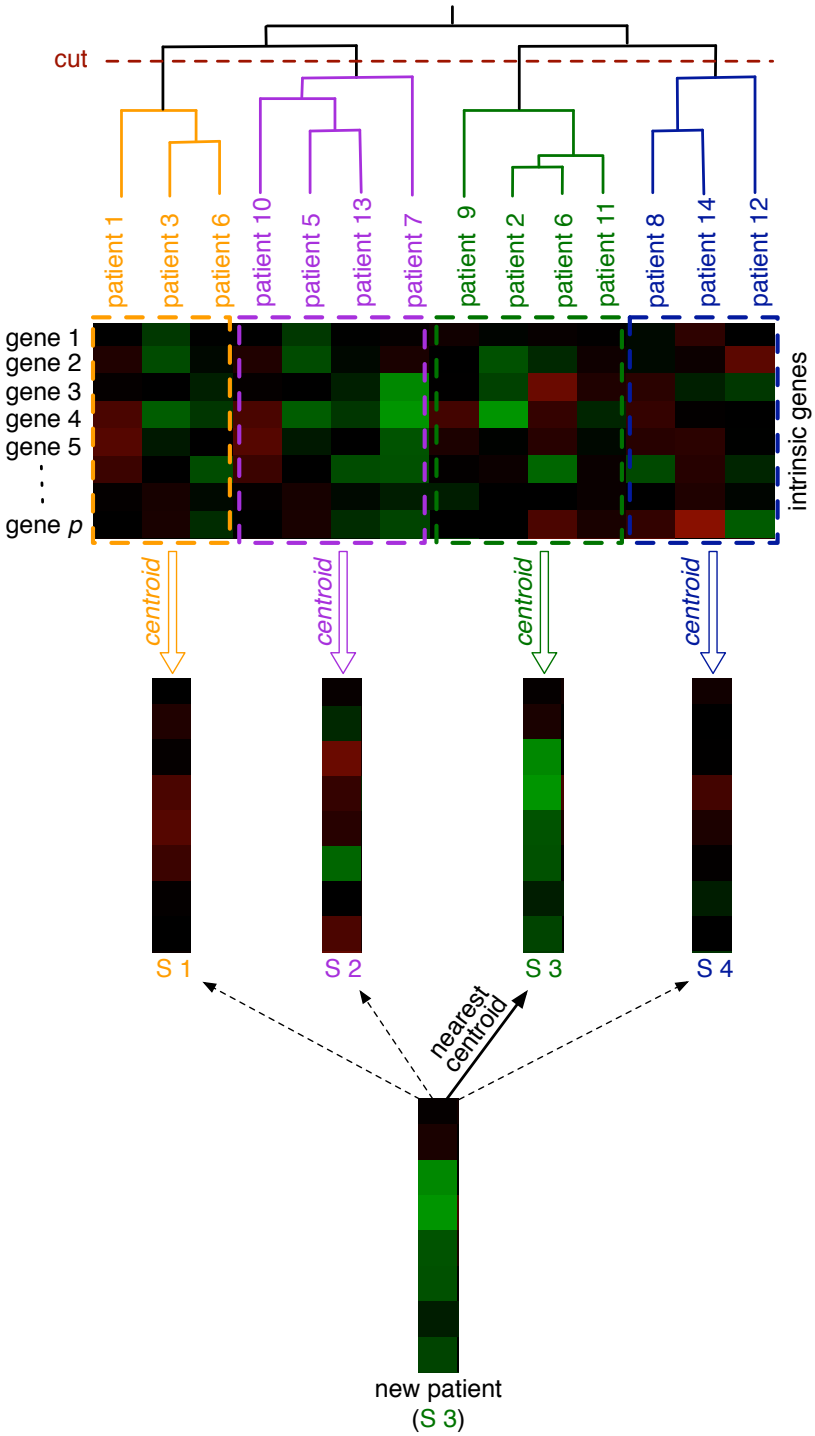


Figure 3.2: Illustration of the method used by Perou et al. to identify breast cancer molecular subtypes. A hierarchical clustering is performed by using the intrinsic gene list to generate a dendrogram of patients' tumors. The dendrogram is then cut to identify the different subtypes (in this case, S1 to S4). A centroid is computed for each subtype. A nearest centroid approach is used to classify a new patient's tumor. In this case, the new tumor is highly correlated with centroid S3, making this the nearest centroid. So the new tumor is predicted to be of the subtype 3.

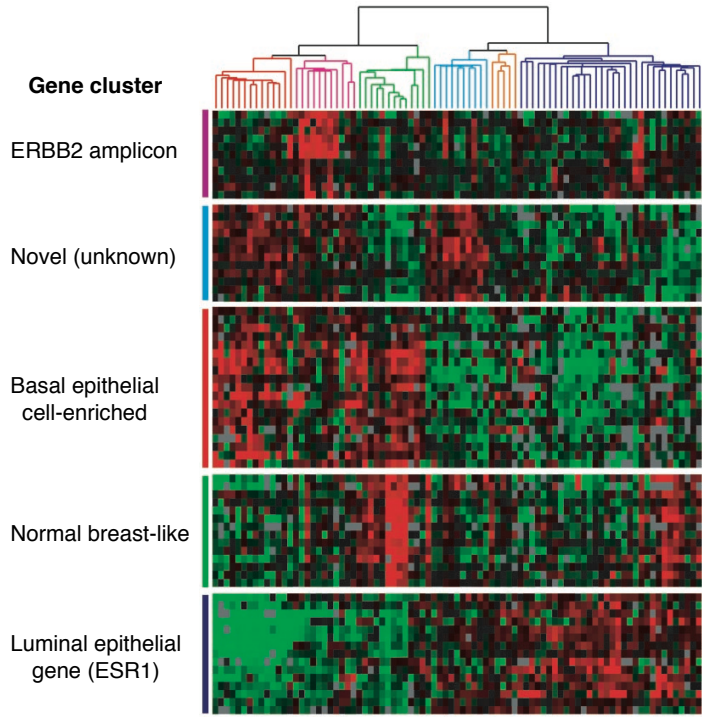


Figure 3.3: Heatmap of the intrinsic genes in [Sorlie et al., 2001]. Sorlie et al. stated that the clustering was mainly driven by the genes related to ER (luminal epithelial gene cluster, ESR1) and HER2 (ERBB2 amplicon gene cluster) signaling pathways. The dendrogram at the top of the heatmap is detailed in Figure 3.4.

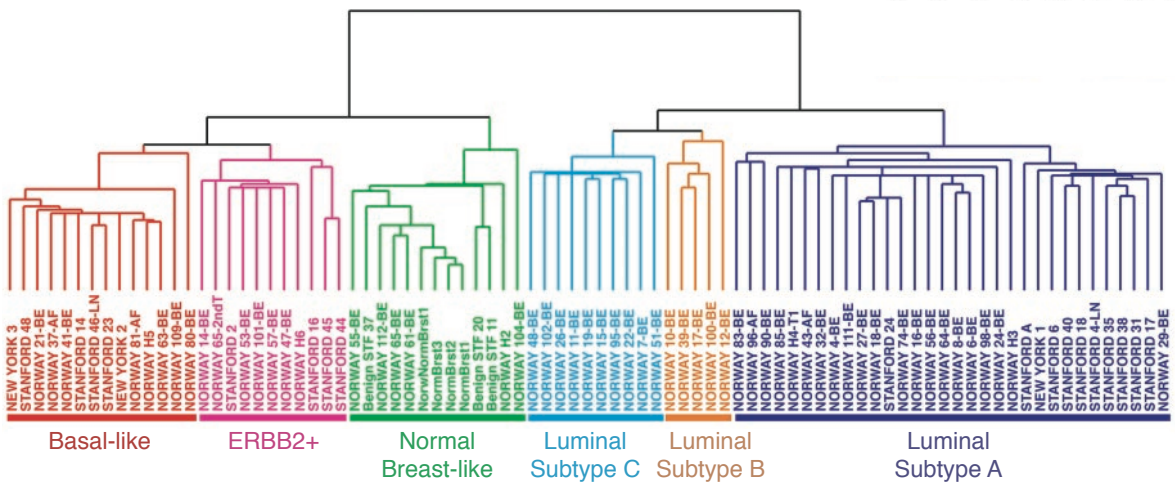


Figure 3.4: Breast cancer molecular subtype identification in [Sorlie et al., 2001]. In this study, Sorlie et al. found six subtypes, namely the basal-like, ERBB2+, normal breast-like, and luminal subtypes A, B and C.

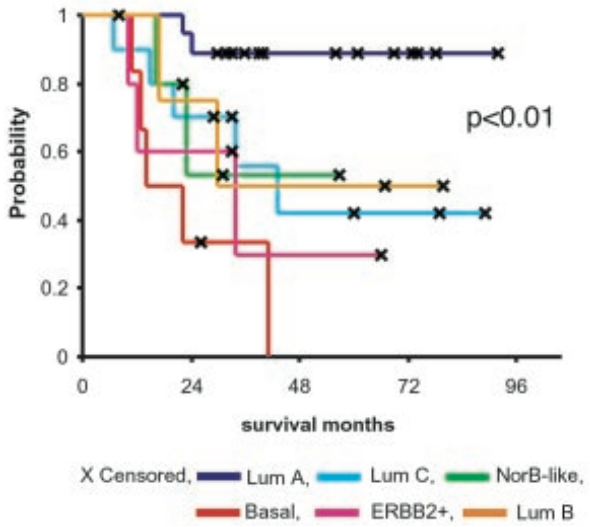


Figure 3.5: Survival curves of the different breast cancer molecular subtypes in [Sorlie et al., 2001].

- The number of clusters (subtypes) that could be reliably identified from the gene expression data was selected in a subjective way (visualization assessment by the analyst). The field lacked of the use of existing statistics to address this issue.
- Similarly, only few studies used existing statistics to validate the robustness of a clustering model, rendering it difficult to assess the reproducibility of the published results.

Concerned by the lack of robustness of these early results, a recent study, published by Kapp et al., showed that only three subtypes could be robustly identified: the ESR1-/ERBB2-, ERBB2+, and ESR1+/ERBB2- subtypes as defined by the expression of a pair of genes, BCMP11 and ABCC11, highly correlated with ESR1 and ERBB2 genes respectively [Kapp et al., 2006]. The authors used a hierarchical clustering method with only a pair of genes in order to yield more robust classification (low feature-to-sample ratio). A measure similar to prediction strength (Section 2.2.4) was used to assess the robustness of the clustering. The robust subtypes these authors found were fairly similar to the previously described subtypes in [Perou et al., 2000; Sorlie et al., 2001, 2003; Sotiriou et al., 2003]. This study addressed most of the issues presented above, but the nearest centroid classifier used in the paper suffers from the same problem as the one in [Sorlie et al., 2003], i.e. it does not provide an accurate estimate of the classification uncertainty.

3.3 Prognostic Gene Signatures

Although we have witnessed in recent decades the development of several prognostic models using histo-pathological information (e.g. NPI or AOL, see Figure 1.2), it remains a challenge to predict which breast cancer patients will suffer a recurrence and who should therefore receive adjuvant therapy.

Great hope was put into the analysis of gene expression data to improve traditional prognostic models. However, the gene expression data generated through microarray technology are complex (see Section 2.1.2), making their analysis challenging, especially in combination with the survival data used in clinical studies. Studies dealing with breast cancer prognostication from gene expression data (supervised learning, see Section 2.1.3.3) are usually confronted to the following issues:

- The development of a prognostic model from gene expression data is prone to overfitting [Everitt, 2002; Hastie et al., 2001]:
  - Gene expressions are noisy, exhibit high correlation between measurements (gene co-expressions) and are usually available for few samples only (high feature-to-sample ratio).
  - Due to the scarcity of biological samples and the cost of the technology, analysts usually lack validation data, making it difficult to honestly assess the performance of the prognostic models.
- The number of gene expressions used in the prognostic model may be large, making the biological interpretation a challenge.
- The presence of censoring requires the use of methods from survival analysis in order to make full use of the information in survival data. These methods are underused in the field of breast cancer prognostication from microarray data.
- The field lacks a thorough performance assessment and comparison framework for prognostic models.

In the following section we will present the state-of-the-art for the extraction of global and local prognostic gene signatures from microarray data. These two types of gene signatures differ from each other in the following way: the global gene signatures are extracted from the whole dataset without considering the presence of different molecular subtypes, whereas the local gene signatures are extracted from specific molecular subtypes present in the whole dataset.

3.3.1 Global Prognostic Gene Signatures

Prognostication In 2002, van't Veer et al. conducted a comprehensive genome-wide assessment of gene expression profiling in order to build a new prognostic model [van't Veer et al., 2002]. Using a set of 78 tumor samples representative of a global population of early (node-negative) breast cancer patients (subset of NKI dataset, see Table 5.1), these investigators identified a set of genes whose expression is associated with patient survival (supervised learning). The authors dichotomized the survival data into a group of patients who had not developed distant metastasis (good prognosis, low-risk) and group of patients who had developed distant metastasis (poor prognosis, high-risk) within the first five years after diagnosis. A feature ranking was performed to sort the genes based on their correlation with this survival-related binary outcome. The authors built a nearest centroid classifier [Dudoit et al., 2002] using the set of the most relevant genes. A cross-validation was performed to identify the best number of relevant genes to include in the nearest centroid classifier, by

optimizing the sensitivity and the specificity of the resulting classifier. The final classifier is composed of 70 prognostic genes and is denoted by GENE70.

Figure 3.6 sketches the heatmap of the genes included in the GENE70 signature with the tumors sorted by their correlation with the good prognosis centroid. We can see that most of the tumors of high-risk patients (patients having a distant metastasis with the first five years after diagnosis, represented by white rectangles at the bottom of the figure), are correctly classified.

The functional annotation for the genes provided insight into the underlying biological mechanism leading to rapid metastases. Indeed, genes involved in cell cycle, invasion and metastasis, angiogenesis, and signal transduction were significantly upregulated in the poor prognosis signature (for example cyclin E2, MCM6, metalloproteinases MMP9 and MP1, RAB6B, PK428, ESM1, and the VEGF receptor FLT1).

The same research group later published a validation study in which they showed the high prognostic value of the GENE70 signature [van de Vijver et al., 2002]. The survival curves of the good and poor prognosis groups were significantly different (logrank test p-value < 0.001, Figure 3.7) and the good prognosis group exhibited a particularly good clinical outcome. Moreover, the authors showed in the original article that the GENE70 signature outperforms traditional clinical guidelines such as St Gallen and NIH.

When this thesis was started, the study by van't Veer et al. was the first to identify a gene signature for breast cancer prognostication. The authors did not use any *a priori* biological knowledge, such as the presence of different molecular subtypes in the dataset, to build their prognostic model. GENE70 is therefore referred to as a global prognostic gene signature.

Although the initial results were promising, these first studies had several major flaws [Dupuy and Simon, 2007]:

- The authors did not make full use of survival data since they dichotomized the survival data with respect to the appearance of a distant metastasis within the first five years after diagnosis. The precise timing of the event occurrence or the censoring is then lost, replaced by a five-year dichotomy. This might lead to poor estimation of the relevance of each gene for survival prediction.
- Due to the small sample size of the study, feature ranking was not performed inside the cross-validation loop. Indeed, the feature ranking was performed using the whole dataset, while the optimization of the signature size was performed in a cross-validation framework. This might lead to an overly optimistic estimation of the performance, as shown in [Ambroise and McLachlan, 2002; Michiels et al., 2005].
- The dataset used in the validation study included a subset of the 78 samples from the initial study. This might also lead to overly optimistic performance estimates [Michiels et al., 2005].
- The datasets used in these studies contained heterogeneously treated patients (mix of untreated patients and patients treated by either chemotherapy or hormonotherapy). Since the treatment should affect the clinical outcome of the patients, the type of treatment might be a confounding factor from a prognostic point of view.

Prediction In this thesis, we focus on the prediction of resistance to tamoxifen, a widely used hormonotherapy for ER-positive breast cancer patients (Section 1.3.2). Tamoxifen sig-

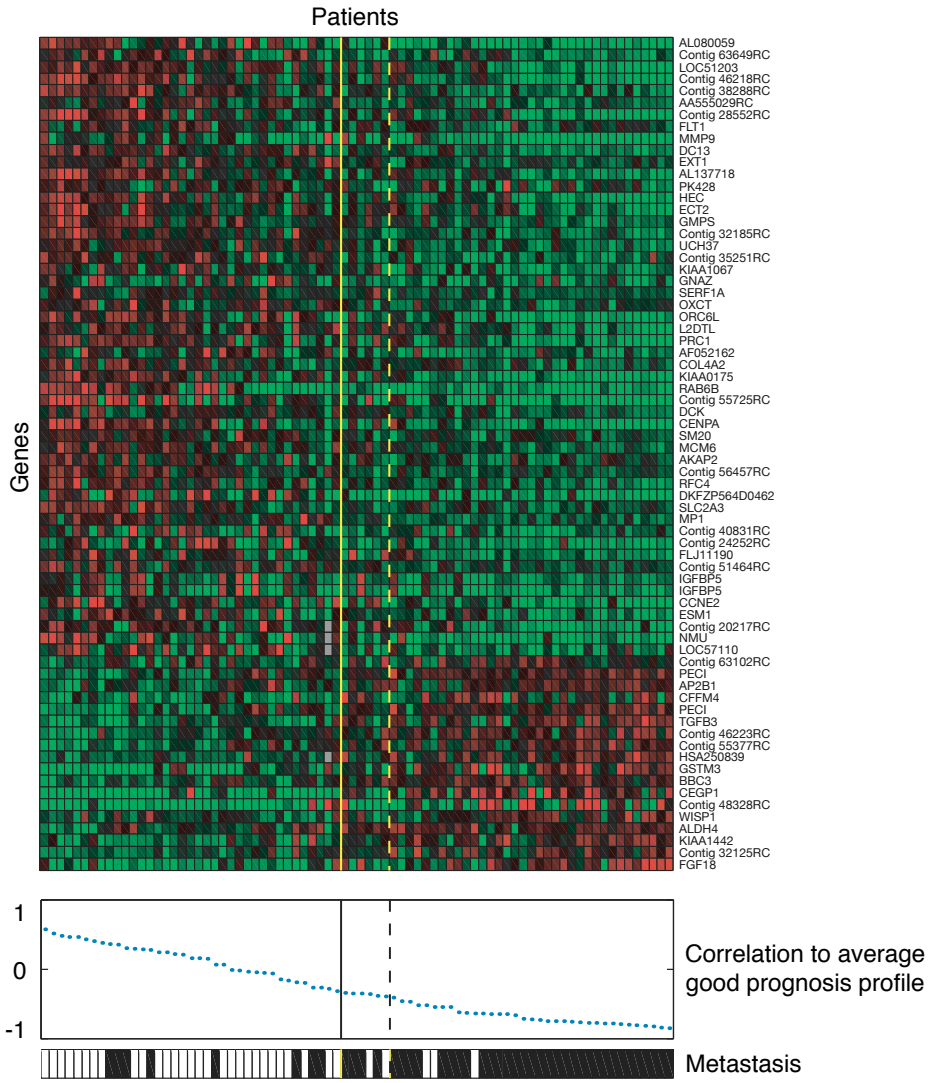
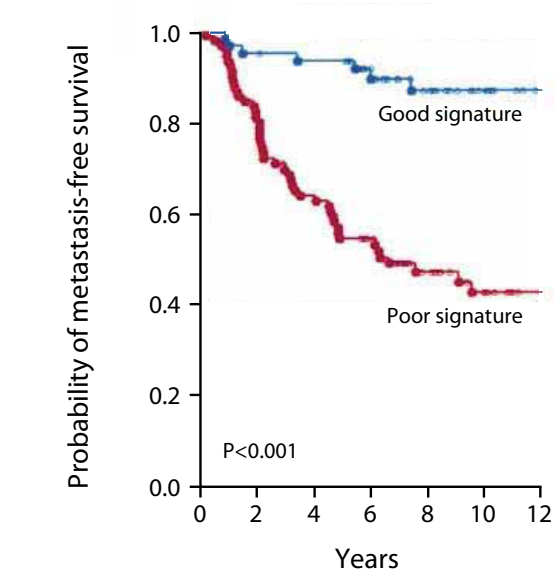


Figure 3.6: Heatmap of the genes included in the GENE70 signature with the tumors sorted by their correlation with the good prognosis centroid in the training set [van't Veer et al., 2002]. The gene names are given in the right side of the heatmap. The solid and dashed yellow lines represent the cutoffs selected to yield best accuracy and sensitivity, respectively. At the bottom of the figure are the risk scores (correlation for each tumor with the good prognosis centroid) and the corresponding risk (black indicates patients who continued to be disease-free for at least five years, white otherwise).



No. at Risk							
Good signature	60	57	54	45	31	22	12
Poor signature	91	72	55	41	26	17	9

Figure 3.7: Survival curves of the poor and good prognosis groups predicted by the GENE70 signature in the population of patients having early (node-negative) breast cancers [van de Vijver et al., 2002].

nificantly reduces tumor recurrence in certain patients with ER-positive breast cancer, but efficient markers predictive of treatment failure have not been identified. In 2004, two studies published a model predictive of resistance to tamoxifen based on gene expression profiling technology, but using different supervised learning approaches [Ma et al., 2004; Paik et al., 2004].

In [Ma et al., 2004], the authors conducted a genome-wide analysis of a set of 60 ER-positive breast cancer patients treated with adjuvant tamoxifen monotherapy (MGH dataset, see Table 5.1). The survival data (distant metastasis free survival, DMFS) were dichotomized in patients who developed distant metastasis with a median time to recurrence of four years (high-risk) and patients who remained disease-free with median follow-up of 10 years (low-risk). A feature ranking was performed to sort the genes based on their differential expression between these two groups (Student t test). A set of nine genes was selected using an arbitrary significance threshold (p-value < 0.001). A ratio of two of these genes, namely HOXB13 and IL17BR, was shown to optimize the area under the ROC curve (Section 2.3.5). The survival curves of the low and high-risk groups predicted by this two-gene ratio are sketched in Figure 3.8. The two groups exhibited significantly different survival (logrank test p-value $<$ 7E-8) in the training set.

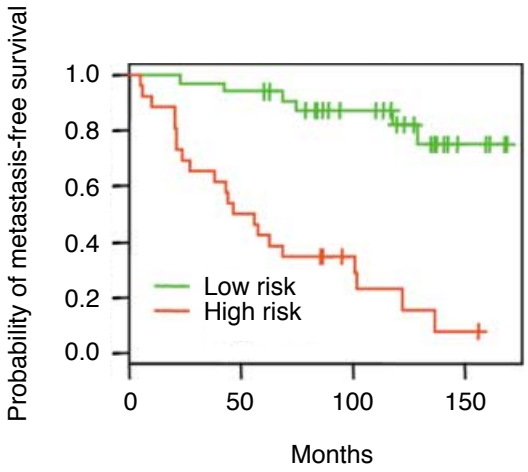


Figure 3.8: Survival curves of the low and high-risk groups predicted by the two-gene ratio in the training set [Ma et al., 2004].

This two-gene ratio was further validated in an independent study of 206 ER-positive breast cancer patients [Goetz et al., 2006]. The survival curves of the risk group predictions (low-risk vs high-risk) computed with the two-gene ratio were significantly different (logrank test p-value < 0.001), although the difference in survival was less impressive (hazard ratio of 2.01) than in the original study.

So the authors concluded that the two-gene ratio (HOXB13:IL17BR) might be useful for identifying patients appropriate for alternative therapeutic regimens in early-stage breast cancer.

In [Paik et al., 2004], the authors used a multistep approach to study resistance to tamoxifen. First, a low-throughput real time reverse transcriptase polymerase chain reaction (RT-PCR) method was developed to perform the gene expression profiling of sections of

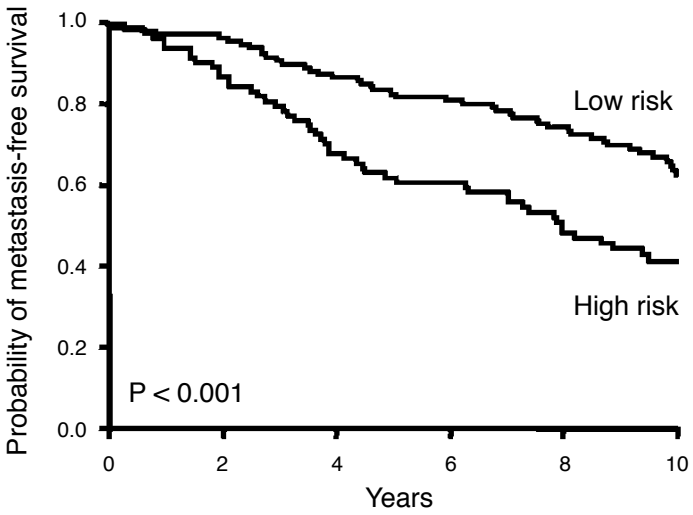
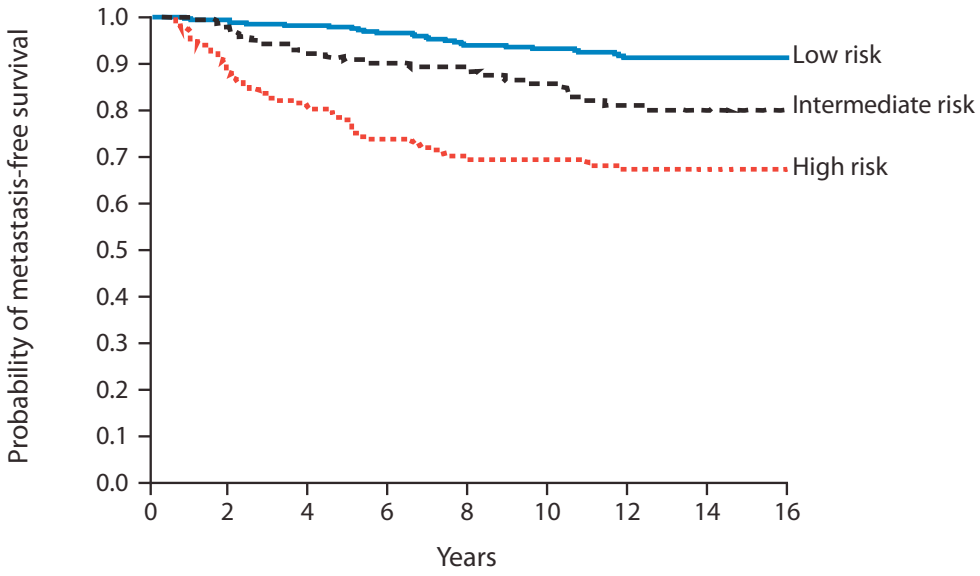


Figure 3.9: Survival curves of the low and high-risk groups predicted by the two-gene ratio in the validation set [Goetz et al., 2006].

fixed paraffin-embedded tumor tissue. Second, the authors selected 250 candidate genes from the published literature, genomic databases, and experiments based on DNA arrays performed on fresh-frozen tissue [Golub et al., 1999; Perou et al., 2000; Sorlie et al., 2001, 2003; van't Veer et al., 2002; van de Vijver et al., 2002]. Third, the authors analyzed data from three independent clinical studies of breast cancer involving a total of 447 patients to test the relation between the expression of the 250 candidate genes and the recurrence of breast cancer. The predictive model fitted to these datasets consisted in the linear combination of the expression of 16 genes involved in proliferation, tumor invasion, ER and HER signaling pathways (see Section 3.1 for details). This set of genes is referred to as ONCOTYPE. Fourth, the authors used this model to predict the resistance to tamoxifen in an independent dataset of 668 ER-positive early (node-negative) breast cancer patients. The risk groups predicted by the ONCOTYPE signature exhibited significantly different survival (logrank test p-value < 0.001), as depicted by the survival curves in Figure 3.10.

Although these initial results were also promising, these studies had several flaws:

- In [Ma et al., 2004], the authors did not make full use of survival data since they dichotomized the survival data with respect to the appearance of a distant metastasis within the first four years of adjuvant tamoxifen therapy. The precise timing of the occurrence of the event or the censoring is then lost, replaced by a four-year dichotomy. This may lead to poor estimation of the relevance of each gene for survival prediction.
- In [Ma et al., 2004], the sample size of the training set was small (60 patients). However, the authors succeeded to validate their results in [Goetz et al., 2006], although the performance in the validation study was poorer (Figure 3.9).
- In [Paik et al., 2004], the authors limited their gene expression profiling to 250 genes extracted on the basis of their relevance in the published literature. Although the resulting predictive model yielded good performance, this limitation has restricted the



No. at Risk									
Low risk	338	328	313	298	276	258	231	170	38
Intermediate risk	149	139	128	116	104	96	80	66	16
High risk	181	154	137	119	105	91	83	63	13

Figure 3.10: Survival curves of the low, intermediate and high-risk groups predicted by the ONCOTYPE signature in the validation set [Paik et al., 2004].

development of new biological insights into the mechanisms responsible for tamoxifen resistance.

3.3.2 Local Prognostic Gene Signatures

Prognostication In 2005, Wang et al. conducted a study similar to [van't Veer et al., 2002; van de Vijver et al., 2002], involving 286 tumor samples from early (node-negative) breast cancer patients (subset of VDX dataset, see Table 5.1) [Wang et al., 2005]. The authors dichotomized the survival data in the same way as [van't Veer et al., 2002]. Wang et al. were the first to propose the development of a prognostic model by dividing the global population of patients into subgroups based on their ER status as defined by immunohistochemistry. A feature ranking was performed for the ER+ and ER- patients separately. For each of the subgroups, risk prediction was computed as a linear combination of univariate Cox's models. A final risk prediction was computed by combining the risk prediction for the two subgroups. The prognostic model was trained on part of the full dataset (115 patients) and validated on the remaining samples (171 patients). The best number of genes to include in the prognostic model was identified by optimizing the sensitivity and the specificity of the prognostic model in the training set. The list of prognostic genes for the ER- and ER+ is denoted by GENE76.

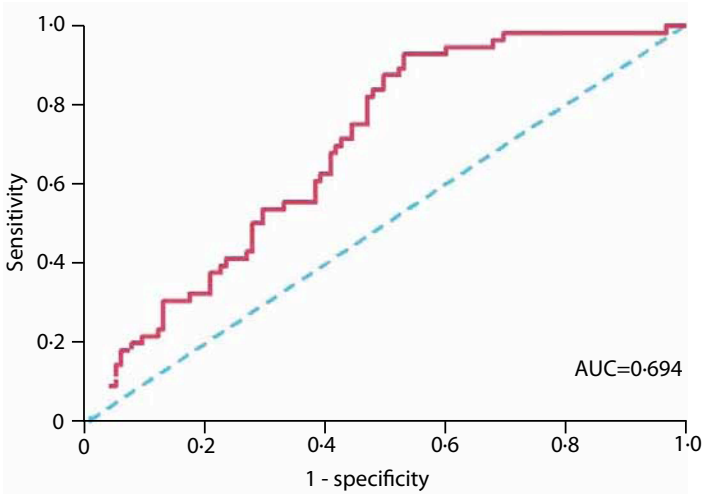
Figure 3.11 (a) shows the ROC curve (Section 2.3.5) of the GENE76 risk group predictions (good vs poor prognosis groups) in the validation set. We can see that the ROC curve is far from the diagonal with an area under the curve (AUC) of 0.694. However, the authors did not provide the significance of such an AUC compared to the null model (diagonal line). The survival curves corresponding to these risk group predictions, are given in Figure 3.11 (b). The difference between the two curves was significant (logrank test p-value < 0.0001) with a hazard ratio of 5.67.

The genes included in the GENE76 signature belong to many functional classes, which suggested that different paths could lead to disease progression. The signature included well-characterized genes and 18 unknown genes. This finding could explain the superior performance of this signature compared with other prognostic factors. Although genes involved in cell death, cell proliferation, and transcriptional regulation were found in both groups of patients stratified by ER status, the 60 genes selected for the ER-positive group and the 16 selected for the ER-negative group had no overlap. This result supported the idea that the extent of heterogeneity and the underlying mechanisms for disease progression could differ for the two ER-based subgroups of breast-cancer patients.

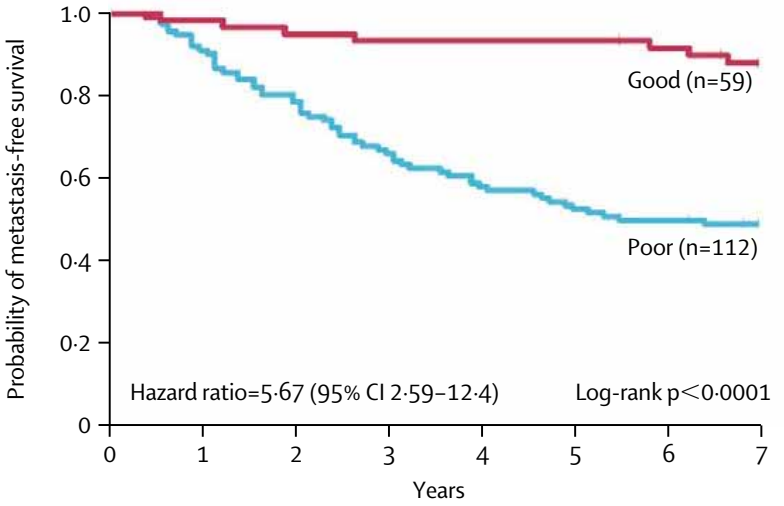
The performance of the GENE76 signature was further validated in an independent study of 180 early (node-negative) breast cancer patients from multiple institutions [Foekens et al., 2006]. The authors reported the survival curves for the good and the poor prognosis groups as predicted by the GENE76 signature (Figure 3.12). The difference between the two curves was significant (logrank test p-value < 0.0001), with an impressive hazard ratio of 6.50.

When beginning with this thesis, the study of Wang et al. was the first to identify a prognostic gene signature for breast cancer taking into account the molecular heterogeneity of the ER phenotype. Although the stratification of the dataset based on the ER status of the tumors (defined by immunohistochemistry) is not as precise as the identification of the molecular subtypes from microarray data, GENE76 can be considered as the first local prognostic gene signature.

Although this prognostic model yielded a promising performance, some issues arose



(a)



Patients at risk

Good signature	59	58	56	55	55	55	53	48
Poor signature	112	103	90	75	66	60	55	52

(b)

Figure 3.11: Performance of the GENE76 risk group predictions (good vs poor prognosis groups) in the validation set [Wang et al., 2005]: (a) ROC curve; (b) Survival curves.

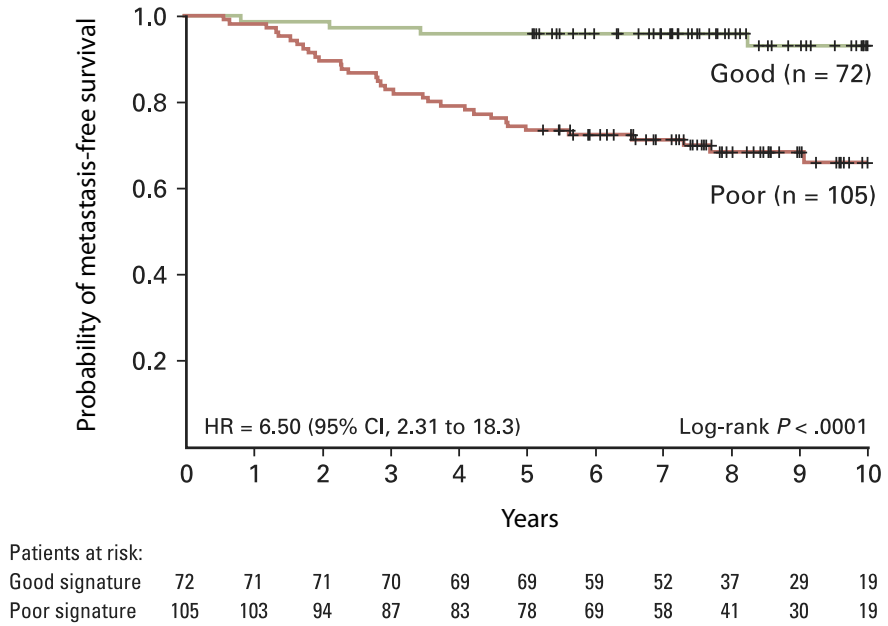


Figure 3.12: Survival curves of the poor and good prognosis groups of patients with node-negative breast cancers [Foekens et al., 2006].

regarding the methodology used:

- The authors considered only two subgroups of patients (ER- and ER+), without taking into account the HER2+ subgroup, shown to be a relevant breast cancer molecular subtype (see Section 3.2).
- Binary IHC evaluation of ER was used to identify the subgroups of patients, leading to hard partitioning of the dataset (no estimation of the uncertainty of this stratification).
- The prognostic model specifically developed for ER- tumors was trained on few samples (35 patients) and yielded poor performance in validation study [Foekens et al., 2006] (data not shown).

3.3.3 Concluding Remarks

Although these initial results for global and local gene signatures were promising, several issues remained open:

- The relationship between traditional histo-pathological parameters and gene expressions. Since such parameters were previously shown to be highly prognostic, it would be interesting to study their molecular basis with the hope to improve their measurement and their prognostic value. In this thesis, we will present the experimental findings of such a study of histological grade in Section 5.2.1.
- The studies related to the prediction of resistance to tamoxifen considered either only few samples [Ma et al., 2004] or few gene expressions [Paik et al., 2004]. There is a

need for a genome-wide study with large sample size to bring new biological insights into tamoxifen resistance and to build a robust predictor. In this thesis, we will present the experimental findings of such a study in Section 5.2.3.

- The relationship between the prognostic genes and the breast cancer molecular subtypes. When work began on this thesis, only few studies attempted to identify the genes and the corresponding biological processes involved in prognosis with respect to the breast cancer molecular subtypes. In this thesis, we will present the experimental findings of such a study in Section 5.4.1.

- The integration of the molecular subtype identification into the development of a prognostic model. Extending the method used in [Wang et al., 2005], it would be interesting to develop a prognostic model integrating an accurate method of subtype identification and the discovery of specific prognostic signatures. In this thesis, we will present the experimental findings of such a study in Section 5.4.2.

3.4 Performance Assessment and Comparison of Prognostic Gene Signatures

Performance assessment and the comparison of prediction models are key steps in microarray data analysis dealing with risk prediction. These steps are often overlooked and may be biased in favor of the risk prediction model presented in the corresponding paper, leading therefore to overly optimistic conclusions. This is particularly true in the field of breast cancer prognostication using microarray data, in which gene signatures have been shown to systematically outperform the traditional clinical tools in an initial paper, these conclusions being rarely confirmed in subsequent validation studies [Eden et al., 2004; Tibshirani and Efron, 2002].

3.4.1 Performance Assessment

Several criteria exist for the performance assessment of risk score and risk group prediction (Section 2.3.5). However, there is no gold standard for the choice of a performance criterion in survival analysis. Most studies dealing with risk prediction in breast cancer use the hazard ratio as computed by Cox's regression (Sections 2.3.4.2 and 2.3.5) or the logrank test for difference between survival curves (Section 2.3.5.2); see [van de Vijver et al., 2002; Wang et al., 2005], to name a few. When beginning this thesis, only few studies used alternative performance criteria such as the concordance index in [Kattan, 2004]. To the best of our knowledge, no study has used several criteria to assess the impact of the choice of the performance criterion on the results of the analysis.

Although hazard ratio is by far the most widely used performance criterion in survival analysis, it suffers from some drawbacks:

- Since the hazard ratio is estimated through Cox's regression, the proportional hazards assumption must hold. The impact of the departure from this assumption is difficult to assess, although it does not seem to affect the hazard ratio estimation dramatically [Cox, 1972] .

- The interpretation of the hazard ratio depends on the scale of the input variable (e.g. risk score), since this ratio represents the difference in risk per unit and per time interval. For instance, let r be a risk lying in $[-1, 1]$; a hazard ratio of 2 means that the hazard of a patient of risk $r = 1$ is twice the hazard of a patient of risk $r = 0$ and four times the hazard of a patient of risk $r = -1$. The scale of the input variable is not always interpretable, rendering difficult the interpretation of the hazard ratio, except in case of ordered class indicators. In the case of binary groups, e.g. low and high-risk groups, the hazard ratio simply represents the difference in risk between any pair of low-risk and high-risk patients.
- The estimation of the hazard ratio is not robust with respect to the presence of outliers.

As for the logrank test, it allows the analyst to assess the significance of the risk group predictions, but it does not allow for quantifying the performance itself. This is a major drawback, especially for performance comparison, as we will see in the next section.

3.4.2 Performance Comparison

Once the performance of the different risk prediction models is assessed, it would be interesting to be able to compare them in order to highlight the potential improvement of any new method or model for breast cancer prognostication for example. The two approaches widely used in the field are the following:

- The multivariate Cox analysis.
- The univariate Cox analysis and naive comparison.

Let r_1 be the risk predictions (risk scores or risk groups) computed by a state-of-the-art method and r_2 be the risk predictions computed by a new method.

3.4.2.1 Multivariate Cox Analysis

One of the comparison procedures widely used to show the superiority of the risk predictions r_2 over r_1 is to fit a multivariate Cox model with r_1 and r_2 as explanatory variables:

$$h(t) = \lambda_0(t) \exp(\beta_1 r_1 + \beta_2 r_2)$$

From this multivariate model, the significance (p-value) of each coefficient β is computed in order to assess the relevance of each risk prediction controlling for the other (Section 2.3.4.2). There are three possible cases:

1. One of the risk predictions is significant, the other is not: for instance, if r_2 is significant and r_1 is not, we can observe that, in most papers, the authors do not hesitate to conclude that r_2 outperforms r_1. In fact, this conclusion is only partly true, depending on the correlation between r_1 and r_2. Let r_1 and r_2 be highly correlated. Since r_2 is slightly more relevant for prediction than r_1, the optimization algorithm used to estimate the coefficients β_1 and β_2 in the multivariate model may estimate a large β_2 and a small (close to zero) β_1. This means that, the two risk predictions being highly correlated, they are not complementary and only r_2 is required to yield good performance in the

dataset under study. However, if the dataset is small, the superiority of r_2 over r_1 may be not generalizable. Moreover, this procedure does not quantify the potential improvement of using r_2 instead of r_1. Now let r_1 and r_2 be poorly correlated. If r_2 is significant and r_1 is not in the multivariate model described earlier, we can reasonably conclude that r_2 yields better performance than r_1. However, the analyst is not able to assess whether there is enough evidence in the data to ensure that the superiority observed in the dataset under study will be generalizable.

2. Both risk predictions are significant: in these settings, the two risk scores should not be highly correlated in order to exhibit a certain complementarity from the prediction point of view. So the use of both risk predictions as explanatory variables in a new model should yield better performance.

3. None are significant: both risk predictions and their combinations do not appear to be relevant from a prediction point of view. A careful univariate analysis (see next section) can confirm the poor performance of each risk prediction.

Using this approach, the analyst attempts to answer two different questions simultaneously: (i) "Does r_2 outperform r_1?"; and (ii) "Are r_2 and r_1 complementary?". As we have seen above, the conclusions may be misleading depending on the correlation between the two risk predictions. Moreover, this approach does not quantify the potential improvement of r_2 over r_1. This is particularly important in breast cancer prognostication using microarray, since microarray technology is expensive and the experiments are difficult to carry out. So the performance gain should be large enough to justify the use of this technology instead of traditional prognostic models.

3.4.2.2 Univariate Cox Analysis and Naive Comparison

This approach requires to fit a univariate Cox model for each risk prediction. So for r_1 and r_2, we have

$$h(t) = \lambda_0(t)\exp(\beta_1 r_1)$$
$$h(t) = \lambda_0(t)\exp(\beta_2 r_2)$$

where the hazard ratios for the risk predictions r_1 and r_2 are equal to $HR_1 = \exp(\beta_1)$ and $HR_2 = \exp(\beta_2)$, respectively.

Unlike the multivariate analysis approach, the univariate analysis of the risk prediction r_1 and r_2 makes it possible to assess the performance of each prediction separately (HR), and a *naive* comparison of these performance estimates can help the analyst show the superiority of some risk prediction over another, e.g. $HR_2 > HR_1$. Such a comparison is referred to as *naive*, since the standard error inherent to each performance estimation is not taken into account. Thus the analyst is unable to assess whether there is enough evidence in the data to ensure that the superiority observed in the dataset under study will be generalizable.

To assess the complementarity of the risk predictions under study, the analyst can build a model combining them and assessing the performance of such a model. If the performance of this model is better than it is for each risk prediction separately, this suggests that the risk predictions are complementary. Again, the standard error of the performance estimation is not taken into account, meaning that the analysist is unable to ensure that the superiority observed in the dataset under study will be generalizable.

3.4.3 Concluding Remarks

As mentioned above, few studies used performance criteria other than the hazard ratio and the logrank test. As we will see in Section 4.4.1, alternative performance criteria presented in Section 2.3.5 have attractive properties that make them good candidates for assessing and comparing performance between risk predictions.

The two usual approaches for performance comparison described above share a common drawback; namely, they do not allow the analyst to assess the significance of the potential superiority of one risk prediction over another or the combination of several risk predictions. In this thesis, we will present in Section 4.4.2 a novel framework for statistical performance comparison to address this issue.

Chapter 4

Methodological Contributions

This chapter details the original methods developed in this thesis. The outline of the chapter is the following. First we present the set of methods used for the identification of global prognostic gene signatures, i.e. signature extraction without taking into account the molecular heterogeneity of the breast tumors. Then we describe our novel clustering model to accurately identify the breast cancer molecular subtypes. Lastly, we present the original methodology allowing for the identification of local prognostic gene signatures, i.e. signature extraction integrating the breast cancer molecular identification.

For each of these topics, the methods are presented using the structure below:

1. Description of the motivations to develop a novel methodology in the context of the state-of-the-art.
2. Description of the method itself.
3. Presentation of the corresponding algorithm if required.
   (a) Step-by step description of the algorithm.
   (b) Discussion of the methods' hyperparamaters and their selection procedure.
4. Description of the pros and cons of the method.

4.1 Identification of Global Prognostic Gene Signatures

In this section, we will present the methods we developed or adapted to extract prognostic gene signatures from the whole dataset, without taking into account the presence of the subtypes (Section 3.2).

Due to the intrinsic complexity of microarray data (see Section 2.1.2), signature extraction is a difficult task prone to overfitting [Everitt, 2002; Hastie et al., 2001]. Signature extraction involves many steps as sketched in Figure 4.1. The steps are the following:

1. Feature transformation: This step allows for reducing the dimensionality of the data in an unsupervised manner (Section 2.1.3.3). We adopt here methods which find a structure in the microarray data and exploit such a structure to select part of the genes or to summarize several gene expressions by a new variable, called a *feature* [Hastie et al., 2001; Webb, 2003].

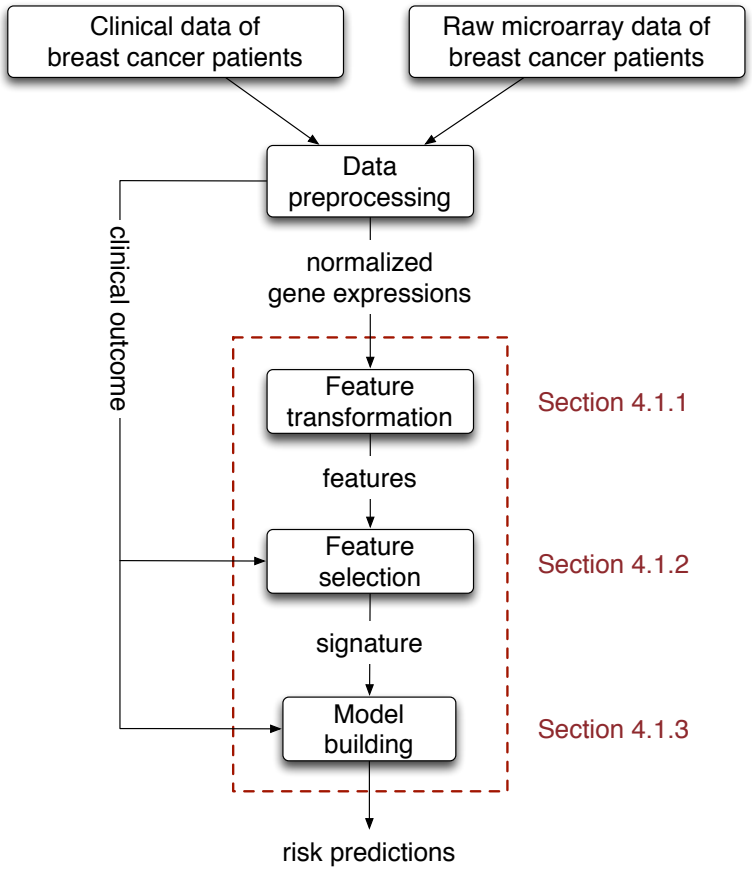


Figure 4.1: Signature extraction methodology. The novel methods developed for the steps delimited by the dashed red box are described in details in the corresponding sections.

2. Feature selection: This step allows for selecting a small set of relevant features in a supervised manner (Section 2.1.3.3), to be used in the prediction model.

3. Model building: This step uses the selected features to build a risk prediction model.

In the following sections, we will present in details the methods for each step of the signature extraction (Figure 4.1). For the preprocessing step, we refer the reader to [Gentleman et al., 2004].

4.1.1 Genome-Wide Feature Transformation

Feature transformation aims at reducing the dimensionality of the input space while retaining most of the information present in the data (Section 2.1.3.2). In the context of the identification of global prognostic gene signatures, we seek to develop a feature transformation method with the following properties:

- The method is able to reduce the dimensionality of genome-wide data, without *a priori* biological knowledge[1].
- The method keeps the new features interpretable from a biological point of view. Indeed, the goal of the identification of prognostic gene signatures is twofold: (i) the building of an efficient prognostic model and (ii) the identification of novel prognostic genes in the signature bringing new biological insights. Feature transformation methods computing complex features should be avoided, since this might dramatically complicate the interpretation of the prognostic signature.
- The method facilitates the computation of the features in datasets using different microarray technologies. Due to the scarcity of frozen tissue samples and the cost of microarray technology, it is often necessary to collect numerous publicly available datasets (Section 5.1) to answer a biomedical question of interest.

Compression and kernel methods (Section 2.1.3.2) do not fulfill these requirements since the computed features are complex, i.e. (non-)linear combinations of large numbers of gene expressions. In contrast, the clustering methods allow to compute features of low complexity, i.e. summaries of highly similar gene expressions. However, some clustering-based feature transformation methods for microarray data require the specification of the number of features to compute [Sheng et al., 2005]. Without *a priori* biological knowledge, the analyst should then tune this hyperparameter, which might be a difficult task. Other feature transformation methods alleviate this issue given that the clustering can be computed without specifying the number of clusters [Eisen et al., 1998; De Smet et al., 2002; Tseng and Wong, 2005]. Such methods include clustering based on hierarchical clustering (Section 2.2.1). However, the biological relevance of the features is not guaranteed since a small number of unknown genes might be clustered together, preventing any interpretation of these clusters. Therefore the clusters of genes should include a number of annotated genes, large enough to be able to link them to known biological pathways.

[1] We will see in Section 4.2.1, how to reduce the dimensionality of the input space when *a priori* biological information are available.

Our approach for genome-wide feature transformation ensures the biological interpretation of the features while facilitating the computation of these features in datasets using different microarray technologies. It consists in identifying clusters of similar genes from the whole set of gene expressions (genome-wide data) in order to summary these clusters by few features (Figure 4.2.):

1. Hierarchical clustering is used to compute the full dendrogram of the gene expressions.
2. This dendrogram is then cut to identify clusters of highly correlated genes.
3. An additional selection is performed to isolate the clusters including a sufficient number of annotated genes in order to facilitate their biological interpretation.
4. Lastly, each cluster of gene expressions is summarized by a feature, therefore reducing drastically the dimensionality of the data.

The procedure is described in Algorithm 2. First, a hierarchical clustering is performed using the function *hclust* (Algorithm 1) in order to identify the nested correlation structure of the gene expressions matrix X. The correlation-based dissimilarity and the average linkage are used in the hierarchical clustering (Section 2.2.1).

Algorithm 2 Genome-wide feature transformation

```
1:  procedure GW.FEATRANSF(X, h, s)
2:      hcl ← hclust(X)
3:      K ← cutree(hcl, h)
4:      I ← {}                                   ▷ cluster of discarded genes
5:      R ← {}                                   ▷ set of prototype genes
6:      for all k_q ∈ K do
7:          if (# annotated genes in k_q) < s then   ▷ use of biological annotations
8:              K ← K \ k_q
9:              I ← {I, k_q}
10:         else
11:             R ← {R, medoid(k_q)}
12:         end if
13:     end for
14:     return (K, R, I)
15: end procedure
```

Once the dendrogram is built, the set K of clusters is identified by cutting the dendrogram at height h using the function *cutree* (Figure 4.2). The hyperparameter h controls for the number of clusters and the collinearity of the gene expressions in the clusters (see next paragraph).

In order to facilitate the biological interpretation of the clusters, those composed of less than s annotated genes[2] are removed from the set K of clusters and the corresponding genes are stored in the cluster I referred to as the cluster of *discarded genes*. Therefore K is composed of clusters with a number of annotated genes large enough to link them to known

[2]The annotations retrieved from public databases, see paragraph on hyperparameters.

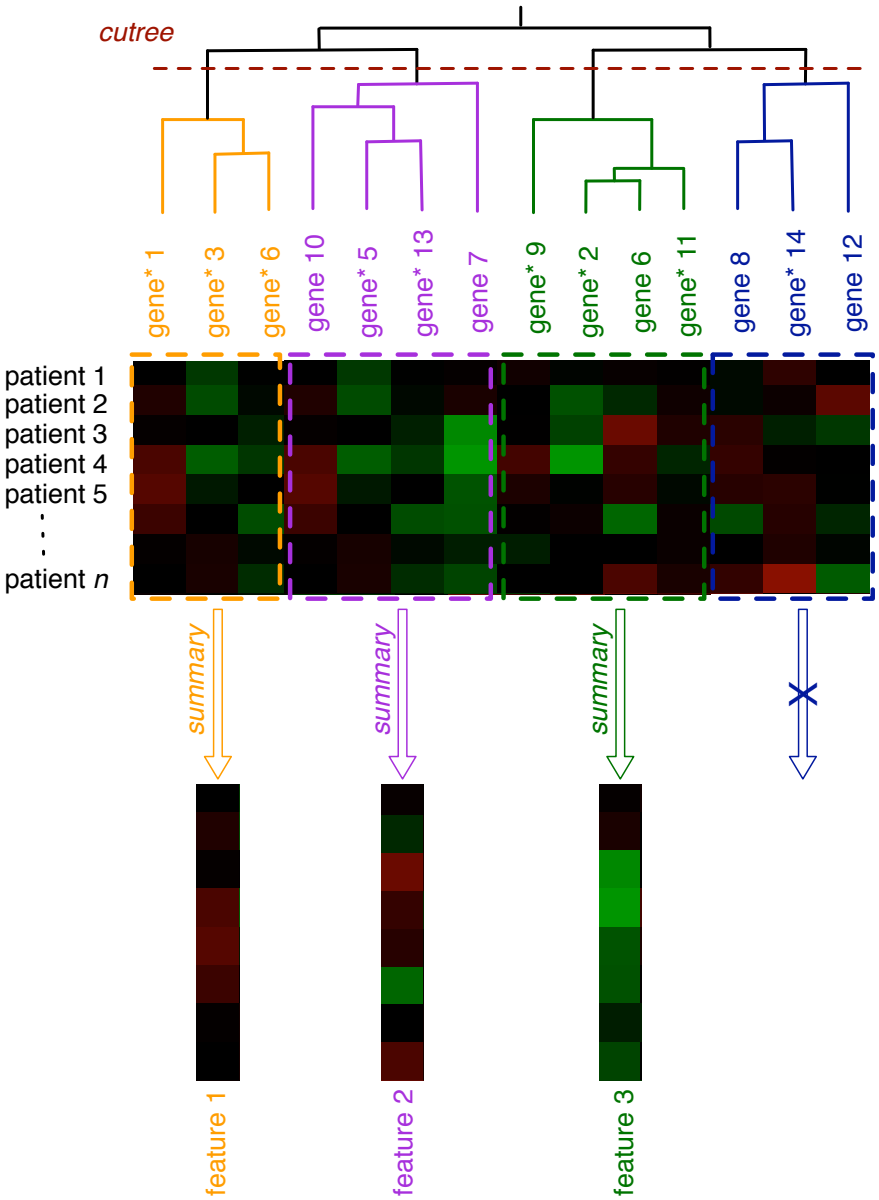


Figure 4.2: Genome-wide feature transformation. The gene are hierarchically clustered in a dendrogram. The dendrogram is cut at a certain height to identify the clusters of similar genes (clusters are differentiated by colors). Clusters that do not include at least 2 annotated genes (the symbol "*" represents the annotated gene) are discarded. Lastly, each remaining cluster is summarized by a new feature.

biological pathways. The cluster *l* contains all the genes which belong to clusters of small size and/or with not enough annotations, possibly empty.

For each cluster of K, a prototype is selected. The prototype of a cluster is its medoid, i.e. a gene selected to represent the cluster of interest and to keep track of the correlation structure [van der Laan et al., 2003]. Indeed, since the correlation-based distance, $1 - |\rho|$ in Equation (2.4), is not influenced by the sign of the correlation estimation, the gene expressions within a cluster may be positively or negatively correlated. We will see in Algorithm 3 how to use this set R of prototype genes to compute the features from the gene expressions within the clusters.

The resulting clustering, composed of the set K of clusters and possibly the cluster *l* of discarded genes, respects the properties presented in Equations (2.2) and (2.3). Algorithm 2 returns the set K of clusters as well as their prototypes R, and the cluster *l* of discarded genes.

To complete the feature transformation, the set K of clusters is summarized by few features in order to reduce the dimensionality of the gene expressions as described in Equation (2.1). To do so, we summarize each cluster by computing a weighted average of the expressions of all the genes within this cluster as described in Algorithm 3. The weights are defined as the signs of the Pearson correlation coefficient ρ of the gene expressions with the prototype r representing the cluster of interest (Algorithm 2). This ensures to have a positive correlation structure within each cluster, each weighted gene expression being positively correlated with the prototype gene. This procedure, from the matrix X of p gene expressions for n patients, returns X', a matrix of features of lower dimensions than the original matrix of gene expressions.

Algorithm 3 Clustering summary

procedure FEATRANSF.SUMMARY(X, K, R)
  for all clusters $k_j \in K$ **do**
    for all $i \in \{1, 2, \dots, n\}$ **do**
      $x'_{ij} \leftarrow \frac{1}{|k_j|} \sum_{m \in k_j} \text{sign}(\rho(x_m, x_{r_j}))x_{im}$ ▷ weighted average
    end for
  end for
  return X' ▷ matrix of features
end procedure

Hyperparameters There are two hyperparameters for the genome-wide approach for feature transformation: the height h at which the dendrogram is cut and the minimum number s of annotated genes in a cluster.

The height h controls for the number of clusters and the collinearity of the gene expressions in the clusters. On the one hand, if h is close to 1, only few clusters composed of large number of genes, will be considered and the pairwise correlation of the gene expressions within the clusters might be low. On the other hand, if h is close to 0, numerous clusters will be considered, each containing few highly correlated gene expressions. This hyperparameter can be fixed by the analyst either (i) by requiring a minimal pairwise correlation of the gene expressions within a cluster to be h; (ii) by optimizing the performance of the clustering

(Section 2.2.4); or (iii) by optimizing through cross-validation [Stone, 1974] the performance of the whole procedure, from the feature transformation to the performance assessment and comparison steps (Figure 4.1). The last procedure is prone to overfitting despite the use of cross-validation techniques due to the high feature-to-sample ratio of the microarray data. Therefore, we do not try to optimize this hyperparameter in our experiments and use instead a value $h = 0.5$ since this value yields good results (Section 5.2.3).

Microarrays may contain numerous probes representing EST, i.e. transcribed sequence from unknown gene, a significant portion of probes being not annotated. This makes the corresponding gene expression not interpretable from a biological point of view. The minimum number of annotated genes s can be fixed by the analyst by looking at the biological database which are publicly available such as:

- NCBI Entrez Gene, `http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene`.
- *e*!Ensemble, `http://www.ensembl.org/index.html`.
- The gene ontology, `http://www.geneontology.org/`.

For instance, the minimum number of genes in the biological pathways present in databases can provide insights into a good number of genes required to be annotated in a cluster to be able to link it to these biological pathways. In our experiments described in Section 5.2.3, we set this hyperparameter to $s = 5$. This facilitated the biological interpretation of the resulting prediction model.

Pros The procedure for feature transformation we developed, has several pros in the framework of microarray data analysis:

- The computation of a feature involves expressions of a sufficient number of annotated genes to facilitate its biological interpretation. The use of tools based on gene ontology [Ashburner et al., 2000] such as Ingenuity Pathway Analysis [Ingenuity Systems] and EASE [Huang et al., 2008], enables to figure out what are the key biological processes in which the clustered genes are involved.
- The features average correlated gene expressions, reducing the variance compared to the original gene expressions. This is beneficial for linear regression (such as Cox regression) as shown in [Park et al., 2007].
- The method facilitates the computation of features in different microarray platforms. Indeed, different microarray platforms are composed of different sets of probes, representing different genes. If few genes of a cluster are absent in the microarray platform of interest, it is still possible to compute the corresponding feature by averaging the expressions of the remaining genes. It may also happen that, in case of very small microarray platforms, all the genes within a cluster are missing, making the computation of the corresponding feature impossible. Fortunately, this is not the case for the datasets used in our experiments (Section 5.1).
- The method takes advantage of the fact that the number of clusters must not be specified in hierarchical clustering to uncover the correlation structure in the original gene expression matrix. Therefore, once the dendrogram is built, the hyperparameters may

be tuned easily. In contrast, the use of a K-means clustering would have required to fit a new clustering model for each number of clusters to identify [Sheng et al., 2005], making the tuning of this hyperparameter computationally burdensome.

- The generalizability of the feature transformation method can be assessed by identifying the clustering (K, I) on a dataset and by validating it (Section 2.2.4) on an independent dataset that will be used for further analysis (feature selection, model building and performance assessment and comparison). This avoids any spurious correlation between the feature transformation step and performance assessment, allowing to obtain dimensionality reduction without increasing the risk of potential overfitting.

Cons

- Unlike compression methods (Section 2.1.3.2), our genome-wide feature transformation method does not allow for estimating the amount of information from the original gene expression matrix that is retained in the smaller matrix of features.

Note that, although the dimensionality of the input data is reduced dramatically, several hundreds of features usually remain, making of feature selection (Section 4.1.2) a necessary step to build a robust predictive model (Section 4.1.3).

4.1.2 Stability-Based Feature Selection

Feature selection allows for selecting a subset of relevant features used to build a prediction model (see Section 4.1.3). In the context of the identification of global prognostic gene signatures, we seek to develop a feature selection method satisfying the following properties:

- The method is not computationally burdensome. The number of possible subsets of features to test is exponential (2^p for p features). Therefore an exhaustive search is not feasible, even after feature transformation, and heuristic approaches are adopted (e.g. ranking or forward/backward feature selection, see [Guyon and Elisseeff, 2003] for a review). The search algorithm used by the feature selection method should be efficient enough to deal with thousands of features.

- The method is robust to overfitting, meaning that the features selected as relevant in the training are also relevant in the independent datasets. Indeed, due to the high feature-to-sample ratio and the noise of microarray data, the feature selection task is challenging and prone to overfitting. Several publications highlighted the lack of stability of the feature selection from the early studies of breast cancer microarray data, no intersection being observed between gene signatures from different datasets [Michiels et al., 2005; Ein-Dor et al., 2005].

- The hyperparameters of the method are easily tuned. Since each feature selection requires the definition of a stopping criterion specifying how many features should be in the gene signature, its tuning should be efficient and intuitive.

Among the three main categories of feature selection methods, namely the filter, wrapper and embedded methods (Section 2.1.3.2), Wessels et al. showed that, in microarray analysis, simple filter methods such as the feature ranking, are particularly adapted thanks

to their low computational cost and their reduced risk of overfitting [Wessels et al., 2002, 2005]. However, the tuning of the stopping criterion, i.e. the number of selected features to include in the signature, is a difficult task given the reduced number of samples and the need of using the same dataset for both feature selection and model building. This is particularly important in clinical studies involving microarray data since this determines the size of the gene signature which is distinctive of the phenomenon under examination.

The traditional approach to identify the best signature size relies on supervised techniques optimizing the performance of the predictive model [Guyon and Elisseeff, 2003]. We will present in this thesis a novel unsupervised technique to achieve this goal.

As supervised approaches, cross-validation techniques have been proposed in literature to select the best number of features [Ambroise and McLachlan, 2002]. Although a cross-validation strategy relies on a multiple fold training and test strategy, it is important to remark that it is still prone to overfitting [Everitt, 2002; Hastie et al., 2001] if it is not kept independent with respect to the model building procedure (see Section 4.1.3). For instance re-using a dataset already employed to select a feature set (e.g. by cross-validation) in order to assess the quality of a predictive model (e.g. again by cross-validation) would return over-optimistic results about the quality of the modeling procedure [Ambroise and McLachlan, 2002]. Another limitation of cross-validation criteria is due to the fact that, like other sampling frameworks (e.g. bootstrap), it generates different subsets of features for each fold or repetition [Michiels et al., 2005; Ein-Dor et al., 2005]. This is particularly annoying in a clinical setting where the variability of the selection reduces the confidence of the doctors in the efficiency of the feature selection procedure.

As an alternative to supervised approaches, Dunne et al. introduced an unsupervised technique, i.e. a technique which does not rely on the performance optimization of the predictive model [Dunne et al., 2002]. This technique attempts to improve the wrapper approach for feature selection by assessing its stability in a sampling framework. Since this initial publication, the stability assessment of feature selection methods received much attention, especially in high dimensional spaces [Kalousis et al., 2005, 2007; Davis et al., 2006; Krizek, 2008]. In this thesis, we develop a stability criterion to identify a good signature size for feature ranking in microarray data analysis. Although the stability of the signature only reduces the variance component of the prediction error (expressed conventionally as a bias/variance sum; [Turney, 1995; Friedman, 1996]), the large amount of noise and the high dimensionality of the input space suggest that this term could be the most important to address in the bias-variance trade-off. The second advantage deriving from the use of stability measures would be a reinforced confidence of doctors in the gene signature outcomes of clinical studies.

4.1.2.1 Feature Ranking

Feature ranking, whose procedure is described in Algorithm 4 consists in two steps:

1. The relevance of each individual feature is assessed according to a univariate scoring function $\mathcal{S}$ supposed to be proportional to the relevance of the feature of interest with respect to the prediction task.

2. All the features are ranked in a decreasing order according to the scores returned by $\mathcal{S}$.

Let y be the survival data, i.e. the time of event t and the censoring indicator c for each patient. Let X be the matrix of p features such that x_j is the j^{th} feature of a patient. Since the prediction task concerns the prognostic of breast cancer patients, the relevance of each feature is defined as its prognostic ability, i.e. the prediction of patients' survival. Each of the performance criteria for survival analysis presented in Section 2.3.5 can be used as scoring function $\mathcal{S}$. Hereafter, we will use the concordance index as scoring function to assess the prognostic relevance of the features since this performance criterion enjoys nice properties in survival analysis (Section 4.4.1). So, the scoring function is defined as

$$\mathcal{S}(x_j, y) = \frac{\sum_{k,l \in \Omega} 1\{x_{kj} > x_{lj}\}}{|\Omega|} \tag{4.1}$$

where x_{kj} and x_{lj} stand for the value of the j^{th} feature of the k^{th} and the l^{th} patient, respectively, and Ω is the set of all the pairs of patients $\{k, l\}$ for whom there is no tie in feature values ($x_{kj} \neq x_{lj}$) and who meet one of the following conditions: (i) both patients k and l experienced an event and time $t_k < t_l$ or (ii) only patient k experienced an event and $t_k < c_l$ where c_l is the censoring time of patient l.

The procedure returns the set of features F containing the k features $x_{(1)}, x_{(2)}, \ldots, x_{(k)}$ having the highest ranking with respect to their prognostic relevance s. This subset of features (signature) will be used to build the prognostic model (see Section 4.1.3).

Algorithm 4 Feature ranking

```
1: procedure FeatRank(X, y, k)
2:     for all j ∈ {1, 2, ..., p} do
3:         s_j ← S(x_j, y)                                  ▷ scoring function
4:     end for
5:     F ← {x_(1), x_(2), ..., x_(k)} where rank(s_(i)) < rank(s_(j)) if i < j
6:     return F                                  ▷ the k most relevant features
7: end procedure
```

Hyperparameter The only hyperparameter of the method is the number k of selected features (called the signature size). In the next section, we present an original unsupervised criterion used to tune this hyperparameter. This novel criterion allows to select the signature size yielding the highest stability.

4.1.2.2 Signature Stability

We present here a criterion assessing the ranking stability with respect to signature size in order to select the size leading to the most stable signature. We adopt an approach based on sampling scheme [Efron, 1981; Good, 2006] to estimate the stability of the feature ranking. The idea is to estimate the distribution of signatures identified through the feature ranking by sampling the original dataset to derive a robust estimate of the signature stability. We illustrate the procedure in Figure 4.3:

1. We sample the original dataset of n patients and select n' patients such that $n' < n$.

2. From these n' patients, we compute the feature ranking to select the k most relevant features (Algorithm 4).

3. The resulting signature is stored and this procedure is performed m times.

4. The stability of the selection of the k most relevant features, denoted by $Stab(k)$, is then estimated from the signatures computed at each of the m sampling steps.

Note that we use sampling *without replacement* (e.g. jackknife or cross-validation; [Efron, 1981]) since the ties artificially created by sampling *with replacement* (e.g. bootstrap; [Efron, 1981]) require complex procedures to be handled properly in survival analysis [Therneau and Grambsch, 2000]. In the experiments described in Chapter 5, we use a sampling procedure which typically samples 90% of the patients to the original dataset.

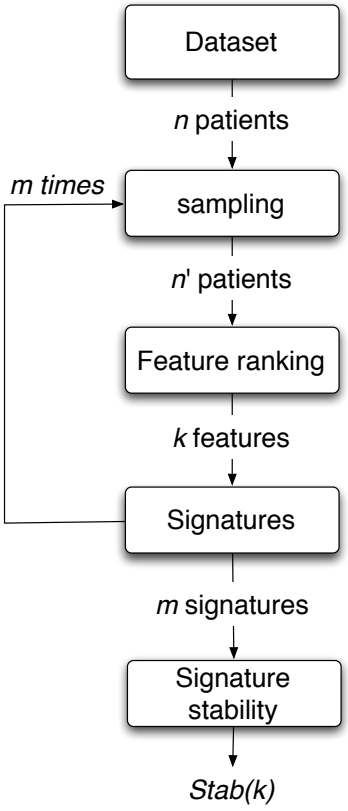


Figure 4.3: Sampling procedure to estimate the stability of a signature of size k selected through feature ranking. The stability of the signature of size k is denoted by $Stab(k)$.

Let illustrate this procedure by an example. Let X be a dataset of $p = 100$ features for $n = 300$ patients. We want to assess the stability of a signature of size $k = 4$ by sampling $m = 1000$ times the original dataset by selecting randomly 90% of the patients each time, so $n' = 270$. For each sampling, the signature including the 4 most relevant features is identified through the feature ranking (Algorithm 4). We can observe in Figure 4.4 that the signature is not always composed of the same 4 features but its composition depends on the sample. Nine features were selected at least once during the sampling procedure. The stability is

proportional to the blue area of the 4 most frequently selected features. In this example, the signature is fairly stable since the red area is small, the top 4 features being selected frequently during the sampling procedure.

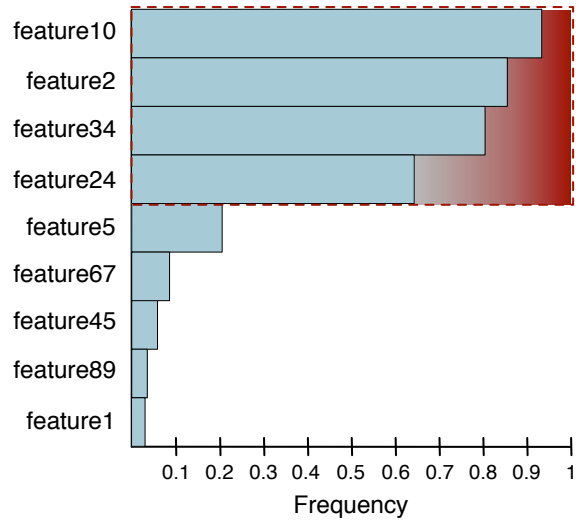


Figure 4.4: Example of stability assessment of a signature composed of 4 features. Smaller is the red area, more stable is the signature.

More formally, let X be the set of p features and $freq(x_j)$ be the number of sampling steps in which a feature $x_j \in X$ has been selected out of m samplings without replacement. The set X is sorted by frequency into the set $\{x_{(1)}, x_{(2)}, \dots, x_{(p)}\}$ where $freq(x_{(i)}) \geq freq(x_{(j)})$ if $i < j$ where $i, j \in \{1, 2, \dots, p\}$. A first measure of stability for a given signature size k is returned by

$$Stab(k) = \frac{\sum_{l=1}^{k} freq(x_{(l)})}{k\, m}$$

This statistic is equal to 1 if the same signature is always selected over sampling steps. In the case of no overlap, $Stab$ is equal to $\frac{1}{m}$ if $k > 0$ and 0 otherwise. However, since the $Stab$ statistic can be made artificially high by simply increasing k, we formulate an adjusted statistic

$$Stab_{adj}(k) = \max\left\{0, Stab(k) - \frac{k}{p}\right\}$$

Thanks to the penalty factor $\frac{k}{p}$, the $Stab_{adj}$ criterion is now equal to 0 for the two extreme cases, i.e. when either no feature or all of them are selected. Moreover, the penalty increases proportionally with the signature size.

Pros The stability-based feature ranking we developed is an intuitive technique which enjoys interesting properties:

- Computational scalability: Feature ranking is computationally efficient since it requires only the computation of the p scores (p being the number of features in the dataset) and the consequent sorting.

- Statistical scalability: Feature ranking, like many filter methods, avoids the estimation of multivariate models to account for the relevance of a set of features. If on the one hand, this exposes the technique to some redundancy (large bias), on the other hand it preserves the approach from overfitting risks (low variance) [Hastie et al., 2001]. This property is particularly appealing in the context of microarray data analysis where the noise is high and the number of features is large, even after feature transformation.
- Hyperparameter: The assessment of signature stability allows for the automatic tuning of the number of selected features while reinforcing confidence for stable gene signatures.

Cons

- Feature ranking usually leads to the selection of a subset of redundant features, i.e. features that could be avoided in the model building procedure without affecting the prediction accuracy [Jakulin and Bratko, 2004]. However, we will see in Section 4.1.3 that the redundancy of the features may be managed at the prediction level by the use of a combination scheme to build the prediction model.
- Feature ranking does not allow to detect complementary features [Wienholt and Sendhoff, 1996; Meyer, 2008]. If several features are irrelevant for prediction individually but highly relevant when combined, feature ranking will not select them since their rank will be large.

4.1.3 Robust Model Building

There exist several alternatives to perform the complex task of building a risk prediction model. Here we enumerate some of the hardest dilemmas to be solved:

- Linear vs non-linear: On the one hand, linear models are simpler, more stable than non-linear models but unable to deal with complex dependencies. On the other hand, the higher complexity of non-linear models reduces the prediction bias at the cost of an increased variance.
- Univariate vs multivariate: Multivariate models deal more effectively with redundancies than univariate ones but demand ill-conditioned and computationally intensive estimation procedures.

The nature of microarray data (large dimensionality even after feature transformation, few samples and high noise) evokes the potential risks of a non-linear and multivariate approach. Actually, we showed in [Haibe-Kains et al., 2008c] that multivariate linear models for risk prediction in breast cancer, although promising in the training set, yielded poor performance in independent datasets. Non-linear risk prediction models, through the use of support vector machine for survival analysis [van Belle et al., 2007], yielded poor performance as well (data not shown). These results highlighted the risk of overfitting with overcomplex risk prediction models. At the same time, a simple univariate model would not be able to account for the multiple interactions underlying the cancer phenomenon. Consequently, there is a demand for a multivariate model which should be able at the same time to return accurate prediction and to avoid instability. An attractive solution to this problem comes from the *additive models*.

Additive models are alternatives to multivariate regression models to deal efficiently with the curse of dimensionality [Stone, 1985; Hastie and Tibshirani, 1990].

Let y be the dependent variable (response) and X be the matrix of k independent variables (predictors). The traditional multivariate linear regression models assume y takes a linear form

$$y = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k \tag{4.2}$$

The additive models estimate an additive approximation to the multivariate regression function such that

$$y = s_1(x_1) + s_2(x_2) + \cdots + s_k(x_k) \tag{4.3}$$

where $s_1(x_1), s_2(x_2), \ldots, s_k(x_k)$ are smooth functions of the independent variables.

The benefits of an additive approximation are twofold. First, since each of the individual additive terms is estimated using a univariate smoother, the curse of dimensionality is avoided, at the cost of not being able to approximate universally. Second, estimates of the individual terms explain how the dependent variable changes with the corresponding independent variables.

The interest of such an additive approach in microarray data analysis lies in the fact that the linear combination of several univariate models returns a model which is simple, yet able to address multivariate tasks. What is less attractive in a context of survival analysis of microarray data is the need of estimating the smooth functions s (backfitting algorithm; [Hastie and Tibshirani, 1990]). A simple method avoiding to fit these functions is provided by combination schemes, commonly used in machine learning literature [Perrone and Cooper, 1993; Kittler et al., 1998], to combine several models in an effective manner. The following section presents the additive solution adopted in the context of the model building for signature extraction.

4.1.3.1 Combination of Models

Several techniques to combine prediction models have been proposed in the literature (see [Kittler et al., 1998] for a review). In this section, we develop an additive prognostic model which combines a set of univariate models to make the risk prediction. Such an additive model takes the form

$$r = \sum_{j=1}^{k} s_j(x_j) \tag{4.4}$$

where $\{x_1, x_2, \ldots, x_k\}$ is the set of selected features (signature of size k, see Section 4.1.2.1), $s_j(x_j)$ is the risk predicted by the j^{th} univariate model (smooth function) and r is the final risk prediction.

Let us start considering the use of univariate Cox models in (4.4). The univariate Cox model using feature x_j as explanatory variable can be written as

$$h(t) = \lambda_0(t) \exp(\beta_j x_j)$$

Since the patients have all the same baseline hazard function $\lambda_0(t)$, the risk of a patient depends only on $r_j = \beta_j x_j$, called the *risk score* in the literature Section 2.3.5. We can thus rewrite (4.4) as

$$r = \sum_{j=1}^{k} \beta_j x_j \tag{4.5}$$

where the β_j are fitted using x_j only (univariate model) in contrast to Equation (4.2).

Due to the complexity of microarray data, we seek to avoid the estimation of the coefficients β_j in order to reduce the risk of overfitting. *Equal weights linear regression* may be an effective answer to this issue. Indeed, several authors showed that under specific conditions, a linear model with equal weights yields robust performance in a validation setting [Wainer, 1976; Green, 1977]. These conditions are the knowledge of the "direction" of the coefficients, i.e. sign(β_j), and the positive inter-correlation of the explanatory variables x_j. Under these conditions, a linear model with equal weights may yield similar or even better performance in independent data than linear model whose weights are estimated from the training data [Wainer, 1976]. Compared to data-driven estimation of the regression coefficients, equal weights regression enjoys two interesting properties. First, it is not sensitive to overfitting since the data are not used to estimate the coefficients. Second, the presence of outliers does not influence the coefficients estimation.

Based on these results, we put equal weight to each univariate model by replacing β_j by $\beta'_j = \frac{1}{k}\,\text{sign}(\beta_j)$. Since only the "direction" of the coefficients β_j should be known to estimate β'_j, as an alternative to the fitting of a univariate Cox model, we can use the scores (concordance indices) computed during the feature selection step (Section 4.1.2.1) such that

$$\beta'_j = \begin{cases} -\frac{1}{k} & \text{if } C\text{-index} < 0.5 \\ +\frac{1}{k} & \text{otherwise} \end{cases}$$

Equation (4.5) then becomes

$$r = \sum_{j=1}^{k} \beta'_j x_j = \frac{1}{k} \sum_{j=1}^{k} \text{sign}(\beta_j) x_j \tag{4.6}$$

which reduces Equation (4.5) to a signed average of the selected features x_j with respect to the "direction" of their prognostic value.

Pros The method for risk prediction model building we propose above is adapted to suvival analysis of microarray data since:

- Additive combination of univariate models allows for addressing multivariate tasks, while exhibiting lower variance than multivariate models.
- Although additive combination of univariate models is less efficient than multivariate modeling to deal with the potential complementarity between features, this class of models is less sensitive to the presence of redundant features, also known as *multicollinearity* in regression. It is typically the case when feature ranking is used to select the relevant features to build the risk prediction model as in this thesis. In this setting,

the prediction accuracy of a multivariate prediction model may dramatically decrease with respect to its complexity, i.e. the number of features used in the model [Myers, 1994]. Additive models are therefore appealing to manage the issue of multicollinearity since only univariate models are used.

- The method used to build the risk prediction model has low computational cost as it avoids the estimation of univariate survival models (e.g. Cox's model) by reusing the scores computed during the feature selection step and by following an *equal weights* approach.

Cons

- Additive models are not able to deal with complementarity of features [Wienholt and Sendhoff, 1996]. This may lead to a lower prediction accuracy of the model.

4.1.4 Concluding Remarks

The original signature extraction methodology presented in this section was designed keeping the following constraints in mind:

- Interpretability: The prognostic gene signature and the resulting predictive model should be interpretable from a biological point of view. To do so, we developed a genome-wide feature transformation method which facilitates the biological interpretation of the resulting features (average of highly correlated annotated gene expressions, see Section 4.1.1) and a simple, yet robust model enabling the interpretation of the contributions of each feature (Section 4.1.3).

- Robustness: The prognostic model should be useable in datasets using different microarray technologies, implying that numerous probes may be missing. In order to respect this constraint, we developed a genome-wide feature transformation method (Section 4.1.1) facilitating the computation of features in various microarray datasets.

- Computational efficiency: The signature extraction should be computationally efficient since we are dealing with thousands of gene expressions. The drastic dimensionality reduction through the genome-wide feature transformation (Section 4.1.1), the fast feature selection through stability-based feature ranking (Section 4.1.2.1) and the robust model building through the combination of univariate models (Section 4.1.3.1) facilitate the extraction of prognostic gene signature from microarray data.

- Accuracy: The prognostic model should yield good performance. Since the variance term in the bias-variance trade-off has a strong impact on the model performance [Ben-Dor et al., 2000; Dudoit et al., 2002; Haibe-Kains et al., 2008c], we designed a signature extraction methodology exhibiting low variance thanks to the summarization of gene expressions in feature transformation (Section 4.1.1), the stable selection of relevant features in stability-based feature selection (Section 4.1.2) and the building of robust (low variance) prognostic model (Section 4.1.3).

Although the method has the desirable properties of interpretability, flexibility, efficiency and robustness, this was made possible at the cost of the capacity to detect complex relations between gene expressions and clinical outcome. Indeed, the stability-based feature

selection prevents to detect complementarity between features while the model building is not able to detect non-linear relations between the gene expressions and the outcome variable. However, numerous authors showed that the intrinsic complexity of microarray data make this task prone to overfitting and difficult to achieve in practice [Ben-Dor et al., 2000; Dudoit et al., 2002; Haibe-Kains et al., 2008c].

4.2 Identification of Breast Cancer Molecular Subtypes

The early microarray studies in breast cancer [Perou et al., 2000; Sorlie et al., 2001, 2003; Sotiriou et al., 2003] highlighted the molecular heterogeneity of breast tumors (Section 3.2). However, the identification of the molecular subtypes appeared to be unstable due to the complexity of microarray data and the small number of samples [Pusztai et al., 2006].

Kapp et al. recently showed that only three subtypes can be robustly identified from microarray data: the ESR1-/ERBB2-, ERBB2+ and ESR1+/ERBB2- subtypes. The resulting clustering model was based on numerous genes correlated with ESR1 and ERBB2 genes and was validated in two independent datasets.

Concerned by the lack of robustness of these initial clustering models, we sought to develop an unsupervised method able to robustly identify the breast cancer molecular subtypes such that:

- The clustering model deals efficiently with different microarray platforms and normalization procedures.
- The clustering model returns an accurate estimate of the classification uncertainty, i.e. for each subtype, the probability for a patient to have a tumor of this subtype.
- The clustering model yields good performance in independent datasets.

The novel method we developed to fulfill these requirements, is composed of two steps (Figure 4.5):

1. Feature transformation: The genome-wide microarray data are transformed into few features quantifying the activity of key biological processes in breast cancer (Section 3.1). This transformation makes possible the identification of breast cancer molecular subtypes in a low dimensional feature space, thus defying the curse of dimensionality particularly relevant in microarray data analysis. In this thesis, we propose a novel feature transformation method which uses a robust estimation of gene co-expression and *a priori* knowledge about the biological processes of interest, to efficiently transform the input space as described in Section 4.2.1.
2. Subtypes identification: we develop a simple, yet robust, model-based clustering in the low dimensional space defined in the previous step, to identify the breast cancer molecular subtypes. This clustering model is described in Section 4.2.2.

4.2.1 Prototype-Based Feature Transformation

The aim of the prototype-based feature transformation is to reduce the dimensionality of the genome-wide microarray data using *a priori* biological knowledge provided by experts. In

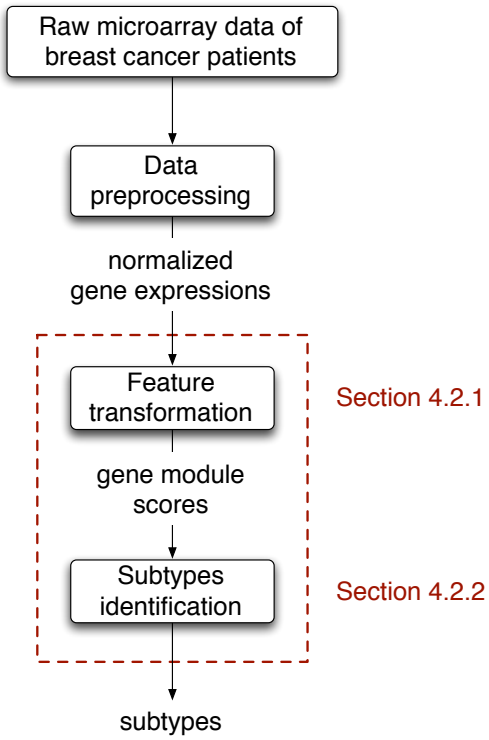


Figure 4.5: Design of the novel unsupervised method used to identify the breast cancer molecular subtypes. The steps delimited by the dashed red box are described in details in the corresponding sections.

particular, we want the new features to quantify the activity of key biological processes in breast cancer (Section 3.1). The features should be *specific* to the biological process they represent, i.e. a feature representative to a biological process should not be also representative to another biological process. This property is referred to hereafter as *biological affinity*[3].

Similarly to the genome-wide feature transformation (Section 4.1.1), the approach adopted for the prototype-based feature transformation is composed of two steps. First, the genes sharing a biological affinity to the same biological process are clustered together. Second, each cluster of genes is summarized by a single feature quantifying the activity at the gene expression level of the corresponding biological process. However, this method differs from the genome-wide feature transformation in the fact that the feature transformation is driven by the *a priori* selection of the key biological processes to study.

To do so, we designed a novel clustering method able to cluster genes with respect to their biological affinity to key biological processes of interest. This method is composed of the following steps:

1. Prototypes selection: In order to select the biological processes to study, the analyst specifies a *prototype gene* for each of them. A prototype gene is a gene known to be representative to the biological process of interest. The choice of prototypes will be described in Section 4.2.1.1.

2. Dissimilarity estimation: For each gene to cluster, the dissimilarities between this gene and the prototype genes are computed to quantify their co-expression[4]. We will introduce in Section 4.2.1.2 a novel dissimilarity estimate to quantify such a gene co-expression.

3. Assignment: Once the dissimilarities are computed, a decision is made to assign the gene to one of the clusters represented by the prototype genes. The framework for this decision making process will be also the subject of a contribution presented in Section 4.2.1.3.

This method is illustrated by the example in Figure 4.6. In this case, the gene expression profiling for three patients was performed. Three prototypes, P1, P2, and P3, are selected by the analyst to represent three biological processes of interest (prototypes selection). In order to cluster gene *j*, the dissimilarities between this gene and each of the prototypes are computed (dissimilarity estimation). Since the dissimilarity (co-expression) between gene *j* and P3 is significantly lower (larger) than with the other prototypes, gene *j* is assigned to the cluster represented by P3 (assignment).

This approach shares many similarities with the *k*-means method [MacQueen, 1967; Hartigan and Wong, 1979], where the number of centers and their position are defined in the prototypes selection step. The novelty of the prototype-based clustering lies in the biology-driven selection of the centers, the dissimilarity estimation and the assignment procedure as we will see in the next sections.

[3] In the original publication [Desmedt et al., 2008], this property was referred to as *specificity*. In this thesis, we use *biological affinity* instead to avoid confusion with the common meaning of *specificity* in statistics.

[4] The dissimilarity is inversely proportional to the level of co-expression (Section 4.1.1). So, two genes exhibiting low dissimilarity are said to be highly co-expressed.

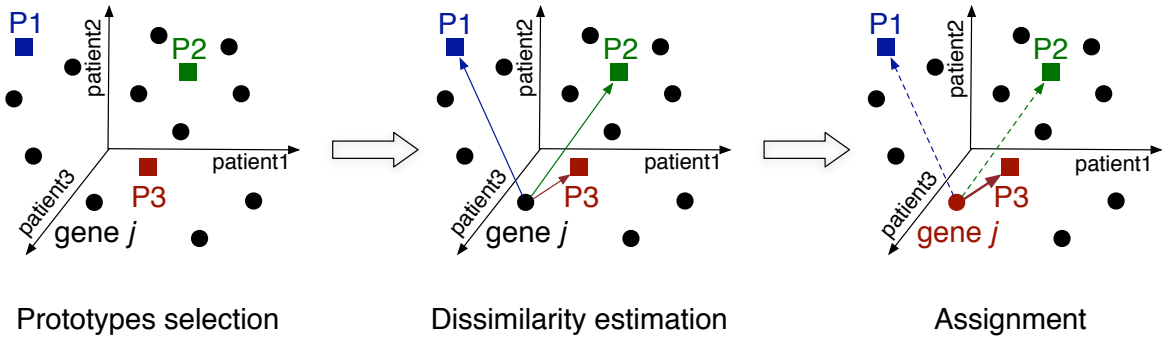


Figure 4.6: Example of prototype-based approach. The three steps of the method are illustrated to finally assign gene j to the cluster represented by prototype P3.

4.2.1.1 Prototypes

In order to drive the feature transformation by some biological processes of interest, the analyst has to specify a prototype gene for each of them. The choice of these prototypes is crucial for the prototype-based feature transformation, since a single gene will be used to represent a whole biological process. Fortunately, in breast cancer, key biological processes were identified these last decades (see [Hanahan and Weinberg, 2000] for a review) and the key genes involved in these biological processes as well. We refer the reader to the Sections 3.1 and 5.3 for a detailed description of the key biological processes in breast cancer and their corresponding prototype genes, respectively.

4.2.1.2 Dissimilarity

We tackle the problem of the reliable estimation of dissimilarity between genes and prototypes from a prediction point of view and use an efficient cross-validation technique for linear models, namely the PRESS (prediction sum of squares) statistic [Allen, 1974].

The PRESS statistics addresses the main issue of cross-validation that is the computationally intensive training-testing procedure for each fold of the cross-validation (Figure A.1 in Appendix A). Indeed, Allen introduces an efficient estimation of the leave-one-out cross-validation (LOOCV) error for linear models as a byproduct of the identification of the linear model coefficients through least squares optimization procedure (see Appendix A for detail)

For the sake of clarity, we define $PRESS(m)$ as the function returning the vector of LOOCV errors of the linear model m through the computation of the PRESS statistic. Let R be the set of prototypes. Similarly to hierarchical clustering (Section 2.2.1), our prototype-based feature transformation requires the estimation of the dissimilarity between a gene and a prototype (called *univariate dissimilarity* hereafter) and the dissimilarity between a gene and a group of prototypes (called *multivariate dissimilarity* hereafter).

Univariate dissimilarity: To estimate the dissimilarity between a gene j and a prototype $r_q \in R$, we calculate

$$d(x_j, x_{r_q}) = \overline{PRESS}(m), m : x_j = \beta_{r_q} x_{r_q} \tag{4.7}$$

where $\overline{PRESS}$ is the mean of LOOCV errors computed by the PRESS statistic. In other words, the dissimilarity between the j^{th} gene and the q^{th} prototype gene is defined as the LOOCV error of the linear model regressing the expression of gene j on the expression of prototype r_q.

Multivariate dissimilarity: Similarly to the linkage in hierarchical clustering (Section 2.2.1), we define the distance between a gene j and a group of prototypes $\{r_1, ..., r_q, ..., r_u\} = R' \subseteq R$ as

$$d(x_j, x_{R'}) = \overline{PRESS}(m), m : x_j = \beta_{r_1} x_{r_1} + \cdots + \beta_{r_q} x_{r_q} + \cdots + \beta_{r_u} x_{r_u} \tag{4.8}$$

where m is the multivariate linear model regressing the expression of gene j on the group of prototypes $\{r_1, ..., r_q, ..., r_u\}$.

Using the dissimilarity measures defined above, we are now able to reliably quantify the co-expression between a gene and each single prototype (univariate dissimilarity) or group of prototypes (multivariate dissimilarity).

4.2.1.3 Assignment

Based on the dissimilarity estimates defined above, we have to define the procedure to assign the genes to the clusters. Traditional methods assign the gene j to the cluster for which the dissimilarity is the lowest [Hastie et al., 2001; Webb, 2003]. However, these methods might be highly unstable since small differences between dissimilarities can be dataset dependent, especially for microarray data given the high feature-to-sample ratio. To address this issue, we develop a new framework for the assignment of the genes to the clusters.

The idea is the following:

1. Set of dissimilarities M: For a gene j, we estimate the dissimilarities to all the single prototypes (univariate dissimilarities) and to all the possible groups of prototypes (multivariate dissimilarities). This set of dissimilarities is denoted by M.

2. Selection of the lowest dissimilarities M': We use the Friedman test [Friedman, 1937] to identify the set of significantly lowest dissimilarities. The Friedman test is a non-parametric statistical test similar to the analysis of variance [Fisher, 1935], often used in model selection to detect differences in errors across multiple statistical models [Birattari et al., 2002]. Therefore, thanks to the Friedman test, we can identify from M the subset M' of the significantly lowest dissimilarities such that the dissimilarities in M' are significantly lower than the dissimilarities in $M \setminus M'$.

3. Biological affinity: Once the set M' is identified, we then evaluate the *biological affinity* of the gene j. The gene j is said to be specific to prototype r_q if the dissimilarity between gene j and the prototype gene r_q is the only univariate dissimilarity in M'. It means that the gene j is significantly more co-expressed to prototype r_q than to all the other single prototypes (univariate dissimilarity) and that the co-expression to prototype r_q is not significantly lower than the co-expression to any groups of prototypes (multivariate dissimilarity).

4. Assignment: The gene j is assigned to a cluster based on its biological affinity. Since the prototype-based feature transformation aims at identifying cluster of genes specifically co-expressed with prototype genes representing biological processes of interest, we only consider clusters represented by single prototypes. So, if a gene j is specific to prototype r_q (biological affinity between gene j and prototype r_q), it is assigned to cluster corresponding to prototype r_q. Otherwise, the gene j is assigned to a cluster of unspecific genes (these genes will not be used for the computation of features).

Figure 4.7 illustrates the use of this method on the example sketched in Figure 4.6. We can see that the genes having low dissimilarities with more than one prototype are not assigned to any cluster since they are not specific (no biological affinity to a single biological process represented by a prototype). The regions delimited by the dashed colored lines are defined by the Friedman test as the regions in which the genes are specific.

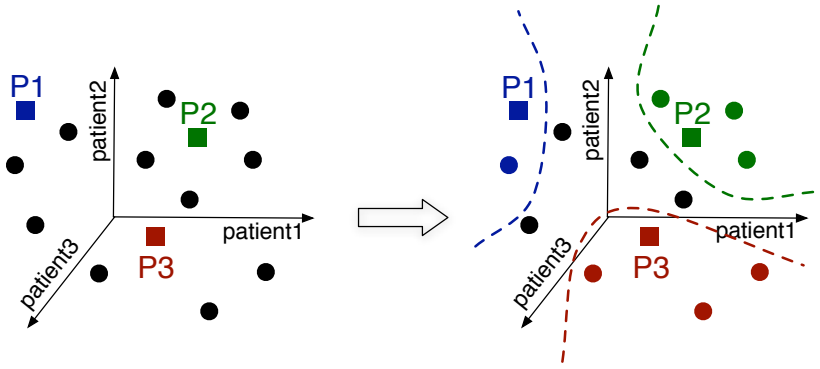


Figure 4.7: Illustration of the prototype-based clustering method on the example sketched in Figure 4.6. Black dots are discarded genes, i.e. genes that are not assigned to any cluster due to the absence of biological affinity to a single biological process represented by one of the prototypes. Colored dots are genes assigned to one of the clusters. The regions delimited by the dashed colored lines are defined as the regions in which the genes are specific.

In practice, the computation of the dissimilarities to all the prototypes and groups of prototypes for each gene is not feasible due to the number of such dissimilarities $\sum_{q=1}^{|R|} \binom{|R|}{q} = 2^{|R|}$ when the number of prototypes $|R|$ is large [Devroye et al., 1997]. For instance, performing the clustering algorithm described above with 7 prototypes for 10,000 genes, requires to compute 1,270,000 dissimilarities. Therefore, we limit the set of dissimilarities to the univariate ones and the lowest multivariate one. However, finding the lowest multivariate dissimilarity for a gene j requires to find the best set of prototypes to predict the expression of gene j in Equation (4.8), which is known to be a NP-complete problem [Davies and Russell, 1994]. Therefore we used a forward feature selection [Kohavi and John, 1997] to identify this best set of prototypes. In particular, we use the orthogonal Gram-Schmidt feature selection since this method was shown to be efficient in combination with the estimation of the PRESS statistic used to compute the dissimilarities [Chen et al., 1989].

The full procedure of the prototype-based clustering is detailed in Algorithm 5. First the set of clusters K is initialized such that each cluster contains a single prototype from the

set of prototypes R. Then, for each gene j from the matrix of gene expressions X, the univariate dissimilarities and the lowest multivariate dissimilarity are computed and put in the set M. These dissimilarities are computed through the *PRESS* function which returns the PRESS statistic of a linear regression model (Section 4.2.1.2). The multivariate dissimilarity is computed through the *orthogonal.Gram.Schmidt* function which returns, from the gene expression matrix X , the set of prototypes and the gene j, the best multivariate linear model to predict the expression of the gene j using the prototypes as explanatory variables.

Algorithm 5 Prototype-based feature transformation

procedure PROTOTYPE.FEATRANSF(X, R, c, e)
  $K \leftarrow \{r_1, r_2, \ldots, r_{|R|}\}$ ▷ clusters initialized with the prototypes
  $I \leftarrow \{\}$ ▷ cluster of discarded genes
  for all gene j **do**
    $M \leftarrow \{\}$ ▷ set of dissimilarities
    for all r_q **do**
      $m \leftarrow x_j = \beta_{r_q} x_{r_q}$ ▷ univariate linear model
      $M \leftarrow \{M, PRESS(m)\}$ ▷ univariate dissimilarity
    end for
    $m \leftarrow orthogonal.Gram.Schmidt(X, R, j)$ ▷ "best" multivariate linear model
    $M \leftarrow \{M, PRESS(m)\}$ ▷ multivariate dissimilarity
    $M' \leftarrow Friedman.test(M, c)$ ▷ set of lowest dissimilarities
    if only one univariate dissimilarity d_q is in $M' \wedge d_q < e$ **then**
      $k_q \leftarrow k_q \cup \{j\}$ ▷ assign gene j to cluster k_q
    else
      $I \leftarrow I \cup \{j\}$ ▷ discard gene j
    end if
  end for
  return (K, R, I)
end procedure

Once the set M of dissimilarities is populated, the Friedman test is used to identify the set M' of significantly lowest dissimilarities. The function $Friedman.test(M, c)$ returns the set of significantly lowest dissimilarities given a critical value c such that the dissimilarities present in M' are significantly lower than models that are not in M' (p-value $< c$) and dissimilarities present in the set are equal (p-value $\geq c$).

If only one univariate dissimilarity d_q is present in the set M' and if this dissimilarity $d_q < e$ where e is the largest acceptable dissimilarity (see the description of the hyperparameters), then the gene j is identified as specific to prototype r_q and is assigned to cluster k_q. Otherwise, the gene j is not specific (the gene j has no biological affinity to any biological processes of interest) and is therefore assigned to the cluster I of unspecific genes.

The resulting clustering, composed of the set of clusters K and the cluster of unspecific genes I, respects the properties presented in Equations (2.2) and (2.3). The algorithm returns the set of clusters K as well as their prototypes r, and the cluster I of discarded genes.

To complete the feature transformation, the set of clusters K, called *gene modules* hereafter, are summarized in order to quantify robustly the activity of biological processes of

interest. As for the genome-wide feature transformation (Section 4.1.1), we summarize each cluster by computing a weighted average of the expressions of all the genes included in this cluster (Algorithm 3). The resulting values are referred to as *gene module scores* in the rest of the thesis.

Hyperparameters There are two hyperparameters for the prototype-based approach feature transformation: the critical value c for the selection of the significantly lowest dissimilarities (Friedman test) and the largest acceptable dissimilarity e for a specific gene.

The critical value $c \in [0, 1]$ allows the analyst to control the error of type I in the Friedman test, i.e. to reject the null hypothesis that two dissimilarities are equal when it is not the case. In our experiments, we use typically a $c = 0.05$ since this value yields good results (Sections 5.3 and 5.4.1).

The largest acceptable dissimilarity $e \in [0, 1]$ allows the analyst to control the strength of the biological affinity. On the one hand, a value of e close to 1 will allow to identify a gene as specific to prototype r_q though its dissimilarity with the prototype r_q is large. On the other hand, a value of e close to 0 will identify genes as specific only if its dissimilarity with the corresponding prototype is very low. In our experiments, we use a $e = 0.95$ since this value leads to an identification of gene modules large enough for the key biological processes in breast cancer (Sections 5.3 and 5.4.1).

Pros The procedure for feature transformation presented above, enjoys interesting properties in the framework of microarray data analysis:

- The features are well defined from a biological point of view since the prototypes are chosen to represent known biological processes.
- The features are computed from specific genes, i.e. genes having a biological affinity to a biological process but not to the others. This ensures the features to be representative to only one biological process, controlling for the others.
- Similarly to quality-based clustering methods [De Smet et al., 2002; Tseng and Wong, 2005], the clustering is not influenced by genes being poorly co-expressed with any other genes. Indeed, these genes are not relevant for clustering and may decrease the robustness of the clustering.
- The features are average of gene expressions, reducing the variance of the measurements. This is beneficial for linear regression (such as Cox regression) as shown in [Park et al., 2007].
- Similarly to the genome-wide feature transformation (Section 4.1.1), the method facilitates the computation of features in different microarray platforms. Indeed, different microarray platforms are composed of different sets of probes, representing different genes. If few genes of a cluster are absent in the microarray platform of interest, it is still possible to compute the corresponding feature by averaging the expressions of the remaining genes. It may also happen that, in case of very small microarray platforms, all the genes within a cluster are missing, making the computation of the corresponding feature impossible. Fortunately, this is not the case for the datasets used in our experiments (Section 5.1).

Cons

- A limitation of the prototype-based feature transformation method lies the fact that only one prototype can be selected to represent each biological process of interest. We will see in Section 5.3 that, using expert knowledge, the key biological processes involved in breast cancer can be resumed by single prototypes. However, this might not be true for other biological processes or other types of cancer where some biological processes should be represented by more than one prototype.

- The procedure used for prototype-based feature transformation is computationally intensive. Indeed, the computation of the dissimilarities, although based on an efficient error estimation technique (PRESS) and reduced to a small set of dissimilarities (only one multivariate dissimilarity), remains computationally intensive in the setting of thousands of genes to cluster.

4.2.2 Subtype Clustering

Thanks to the prototype-based feature transformation method presented above, we are now able to represent the patients in a low dimensional space defined by gene module scores quantifying the activity of the biological processes of interest. In order to identify the breast cancer molecular subtypes, we require the clustering model (i) to return accurate estimate of the classification uncertainty, i.e. the probabilities for a tumor to belong to each of the subtypes; and (ii) to be easily applicable to new data. Note that, unlike the prototype-based feature clustering which clusters the genes, the clustering model described below groups the tumors/patients with respect to their subtype (Figure 4.8).

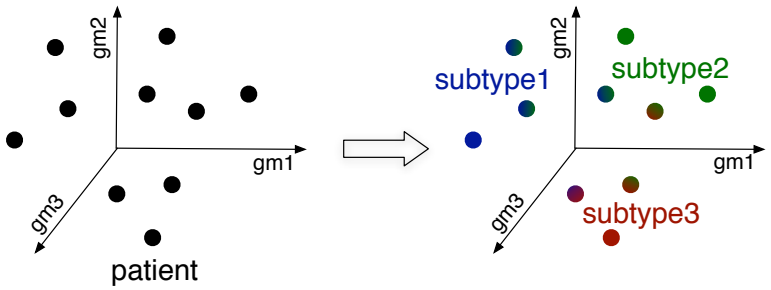


Figure 4.8: Illustration of the subtype clustering. Left side: Representation of the patients' tumors in the low dimensional space defined by the three gene module scores gm1, gm2 and gm3 computed using the prototype-based feature transformation. Right side: Clustering of the patients' tumors according to their subtype (in this case, three subtypes of different colors).

To fulfill these requirements, we develop a model-based clustering that is a mixture of Gaussians (Section 2.2.2) in a low dimensional space. The input space is defined by few gene module scores computed through the prototype-based feature transformation presented above.

Let $X_{n\times p}$ be the matrix of p gene module scores for n patients and x_i be the profile of the

i^{th} patient. A mixture of Gaussians model (Section 2.2.2) can be written as

$$\Pr(x_i) = \sum_{r=1}^{u} \pi_r \mathcal{N}(x_i; \mu_r, \Sigma_r) \tag{4.9}$$

where u is the number of Gaussians, π_r is the prior probability of x_i to be generated by the r^{th} Gaussian $\mathcal{N}(x_i; \mu_r, \Sigma_r)$ of mean μ_r and covariance matrix Σ_r.

From Equation (2.7), we define the probabilities to belong to each subtype r as

$$\Pr(r|x_i) = \frac{\pi_r \mathcal{N}(x_i; \mu_r, \Sigma_r)}{\sum_{s=1}^{u} \pi_s \mathcal{N}(x_i; \mu_s, \Sigma_s)} \tag{4.10}$$

So, $\Pr(r|x_i)$ is the probability that the patient having the profile x_i has a breast tumor of subtype r.

In order to reduce the complexity of the mixture of Gaussians, and so the risk of overfitting [Bishop, 1994; Yeung et al., 2001], we constraint the covariance matrices to be diagonal and equal across the Gaussians such that

$$\Sigma_r = \begin{bmatrix} \sigma_1^2 & 0 & 0 & 0 & 0 \\ 0 & \ddots & 0 & 0 & 0 \\ 0 & 0 & \sigma_q^2 & 0 & 0 \\ 0 & 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & 0 & \sigma_p^2 \end{bmatrix}, \forall r \in \{1, 2, ..., u\} \tag{4.11}$$

It remains to estimate the parameters of the mixture of Gaussians: (i) the number u of Gaussians and (ii) the parameters of these Gaussians, i.e. the means μ and the matrix of covariance Σ. To do so, we adopted (i) the BIC criterion to identify the most likely number of Gaussians (Section 2.2.2.1) and (ii) the EM algorithm (Section 2.2.2; [Dempster et al., 1977]) to estimate the parameters of the Gaussians.

Pros

- The low dimensional input space, defined by the gene module scores computed by the prototype-based feature transformation, yields a low feature-to-sample ratio, increasing the robustness of the clustering model.
- The low dimensional input space facilitates the representation of the results. In the case of breast cancer molecular subtypes, the patient are represented in a 2D plot (see Section 5.3 for details).
- Unlike the hierarchical clustering method used in the initial publications [Perou et al., 2000; Sotiriou et al., 2003] which returns a hard partition of the dataset (Section 2.2.1), the mixture of Gaussians allows for a soft partitioning of the dataset (Section 2.2.2) through the straightforward estimation of the posterior probability to belong to each subtype (classification uncertainty).
- Unlike the hierarchical clustering used in the initial publications [Perou et al., 2000; Sotiriou et al., 2003], this model-based clustering can be easily used to predict the subtype of a new tumor given its gene expression profile.

Cons

- Although the use of few features, defined by the gene module scores, is beneficial for the robustness of the method (low feature-to-sample ratio), it also prevents us to find any new subtypes based on unknown sets of genes relevant for breast cancer molecular subtypes. However, Kapp et al. showed in a genome-wide study without *a priori* selection of genes for the clustering, that robust subtypes were only identified using genes associated with two well-known biological processes, namely ER and HER2 signaling pathways (Section 3.1) [Kapp et al., 2006]. We will see in Section 5.3, that the use of the gene module scores representative to these two biological processes, in our subtype clustering model yields robust identification of the breast cancer molecular subtypes.

4.2.3 Concluding Remarks

The novel method we developed for the breast cancer molecular subtypes identification uses few biologically meaningful features and a model-based clustering of low complexity. This method has several advantages, the most important being the robustness thanks to the low feature-to-sample ratio and the accurate estimation of the classification uncertainty.

It is worth to note that we did not attempt to find new breast cancer molecular subtypes from scratch, but we used instead the results of previous publications, especially [Kapp et al., 2006], to build a robust subtype clustering model. Although this prevents to find new subtypes and to bring new biological insights, the use of some *a priori* biological knowledge about the relevant biological processes involved in breast cancer biology enables the development of a simple, yet robust, method. The robustness of our method will be validated in Section 5.3.

4.3 Identification of Local Prognostic Gene Signatures

The global approach described in Section 4.1 aims at extracting a gene signature from a global population of patients. Such a global risk prediction model makes the assumption that the relationship between the gene expressions and the risk of a patient can be described by an analytical function over the whole domain of the input space. Moreover, it solves the problem of learning by extracting the hypothesis which is expected to approximate the best the whole data distribution. Given that breast cancer, in addition to being a clinically heterogeneous disease, is also molecularly heterogeneous (Section 3.2), global risk prediction is a challenge since the risk prediction model should deal with this heterogeneity to reduce prediction error.

The *divide-and-conquer* paradigm is an attractive alternative to the global risk prediction approach. It originates from the idea of relaxing the global modeling assumption. It attacks a complex problem by dividing it into simpler problems whose solutions can be combined to yield a solution to the original problem. This principle presents two main advantages. First, simpler problems can be solved by simpler estimation techniques such as the adoption of linear models. Second, the model can better fit the properties of the available dataset since the combination of local linear models can achieve non-linear modeling of the data.

In this section, we will present the use of the divide-and-conquer paradigm for *modular modeling*. Modular modeling techniques replace the global risk prediction model with a modular architecture where the modules cover different parts of the input space. Radial Basis Functions [Moody and Darken, 1989], Local Model Networks [Murray-Smith and Johansen, 1997] or Classification and Regression Trees [Breiman et al., 1984] are well-known examples of this approach. In this thesis we propose an original method based on a modular modeling approach to improve breast cancer prognostication (Figure 4.9). In this case, the modules are defined as the molecular subtypes and the local prognostic signatures are extracted specifically for each subtype in order to combine these risk predictions into a global risk prediction for each patient.

The outline of this section is as follows: Section 4.3.1 introduces the modular modeling approach, Section 4.3.2 defines the modules and Section 4.3.3 describes the local risk prediction models, especially the identification of prognostic gene signatures in this framework.

4.3.1 Modular Modeling Approach

The approach for modular modeling used in this thesis is derived from the Local Model Networks (LMN), first introduced in [Johansen and Foss, 1993]. A LMN is a set of different models which are weighted according to the input (Figure 4.10). Each model is used in parallel, the output being multiplied with a basis function and summed to give the local model network output. Essentially the basis functions control the smoothness of the transition as the operating point moves from one local model to another. Indeed each local model has a validity region where the model is most active.

The smooth combination provided by the LMN formalism enables non-linear models on the basis of simpler modules. See the example in Figure 4.11 which allows the combination in a two dimensional input space of three local linear models whose validity regions are represented by Gaussian basis functions.

The general form of a LMN is

$$y = \sum_{j=1}^{m} \rho_j(x, \theta_j) h_j(x, \alpha_j) \tag{4.12}$$

where x and y stand for gene expressions and survival data respectively, $\rho_j(x, \theta_j)$ is the j^{th} basis function depending on x and the parameters θ_j, $h_j(x, \alpha_j)$ is the j^{th} local risk prediction model depending on x and the parameters α_j. The ρ_j in Equation (4.12) are constrained to satisfy

$$\sum_{j=1}^{m} \rho_j(x, \theta_j) = 1 \tag{4.13}$$

This means that the basis functions form a *partition of unity* [Murray-Smith, 1994]. This ensures that in every point of the input space, the prediction is a weighted average of the local models h_j.

Note that the LMN method adopted in this thesis for the identification of local prognostic gene signatures, has a lot in common with the *mixture of experts* [Jacobs et al., 1991] and its extension, the *hierarchical mixture of experts* [Jordan and Jacobs, 1994]. For instance, in

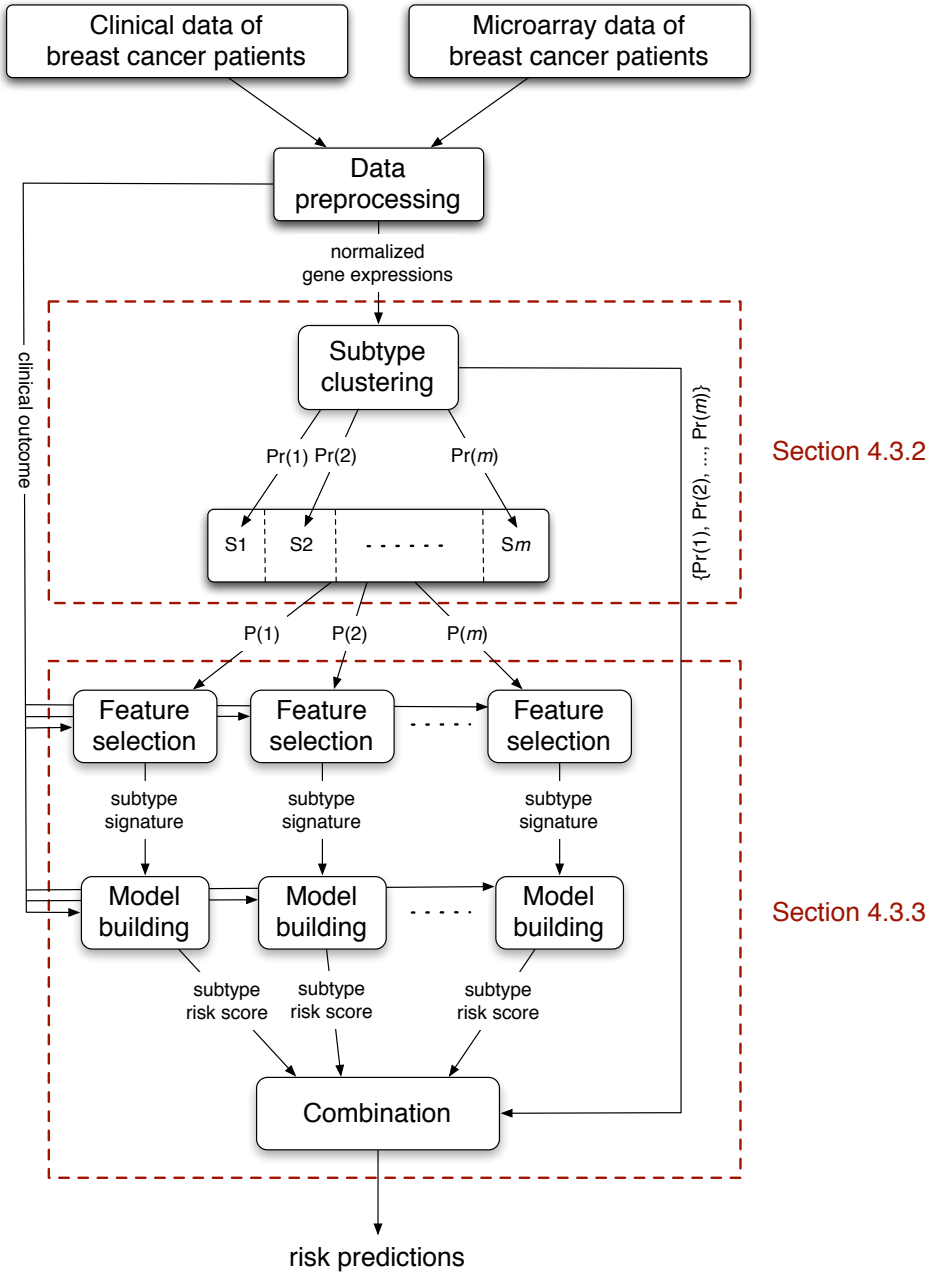


Figure 4.9: Design of the original modular modeling methodology which consists in dividing the global population of patients' tumors with respect to the m molecular subtypes (with probabilities $\Pr(1), \Pr(2), \ldots, \Pr(m)$). Then the feature selection (identification of prognostic genes in the *subtype signatures*) and the building of the risk prediction model are performed for each subtype. Lastly, the resulting local risk predictions (*subtype risk scores*) are combined to compute the global risk prediction of the patients (*risk predictions*). The novel methods developed for the steps delimited by the dashed red boxes are described in details in the corresponding sections.

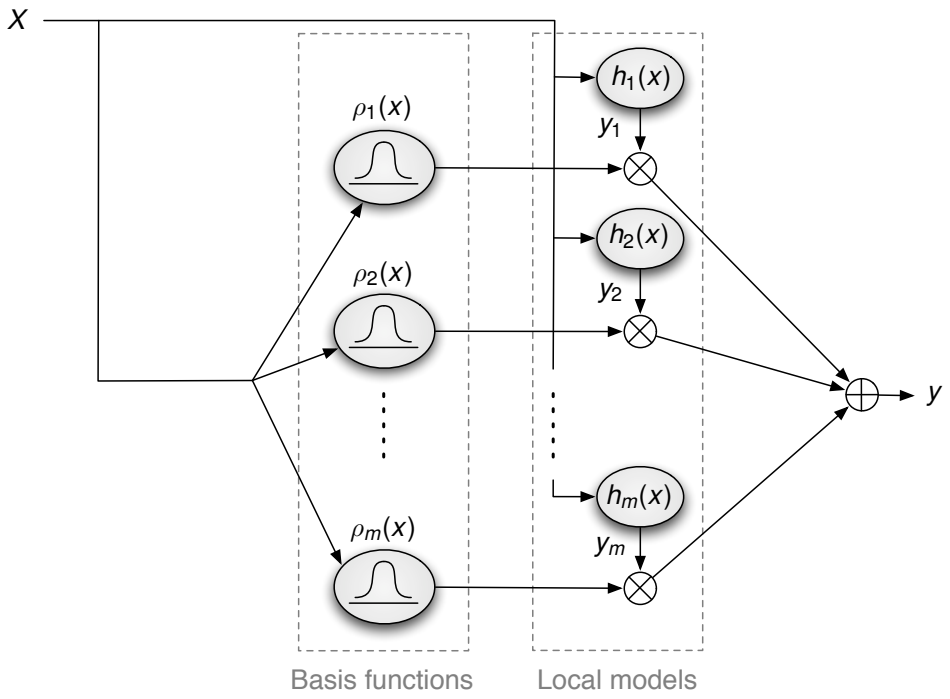


Figure 4.10: General form of Local Model Networks. The input data are denoted by X, basis functions by ρ, local models by h and output data by y.

both methods, the prediction is a weighted combination of the predictions of the local models or experts [Bontempi, 1999].

In the following sections, we will present the identification of the modules through the basis functions ρ and the signature extraction method used to build the local models and the corresponding local prognostic gene signatures.

4.3.2 Modules

In order to apply the modular modeling approach, we need to partition the input space in modules. These modules are defined as the breast cancer molecular subtypes described in the literature (Section 1.1.1 and 3.2). The corresponding basis functions ρ are defined as the probability density functions to belong to each of these subtypes. These probabilities are estimated through the novel subtype clustering model presented in Section 4.2.

4.3.2.1 Breast Cancer Molecular Subtypes

From Equation (4.9), we define the basis functions of Equation (4.12) as

$$\rho_j(x) = \frac{\pi_j \mathcal{N}(x; \mu_j, \Sigma_j)}{\sum_{l=1}^{m} \pi_l \mathcal{N}(x; \mu_l, \Sigma_l)} \tag{4.14}$$

where π_j is the prior probability of x to be generated by the j^{th} Gaussian $\mathcal{N}(x; \mu_j, \Sigma_j)$ of mean μ_j and covariance matrix Σ_j. In this case, the set θ_j of parameters of the j^{th} basis function is

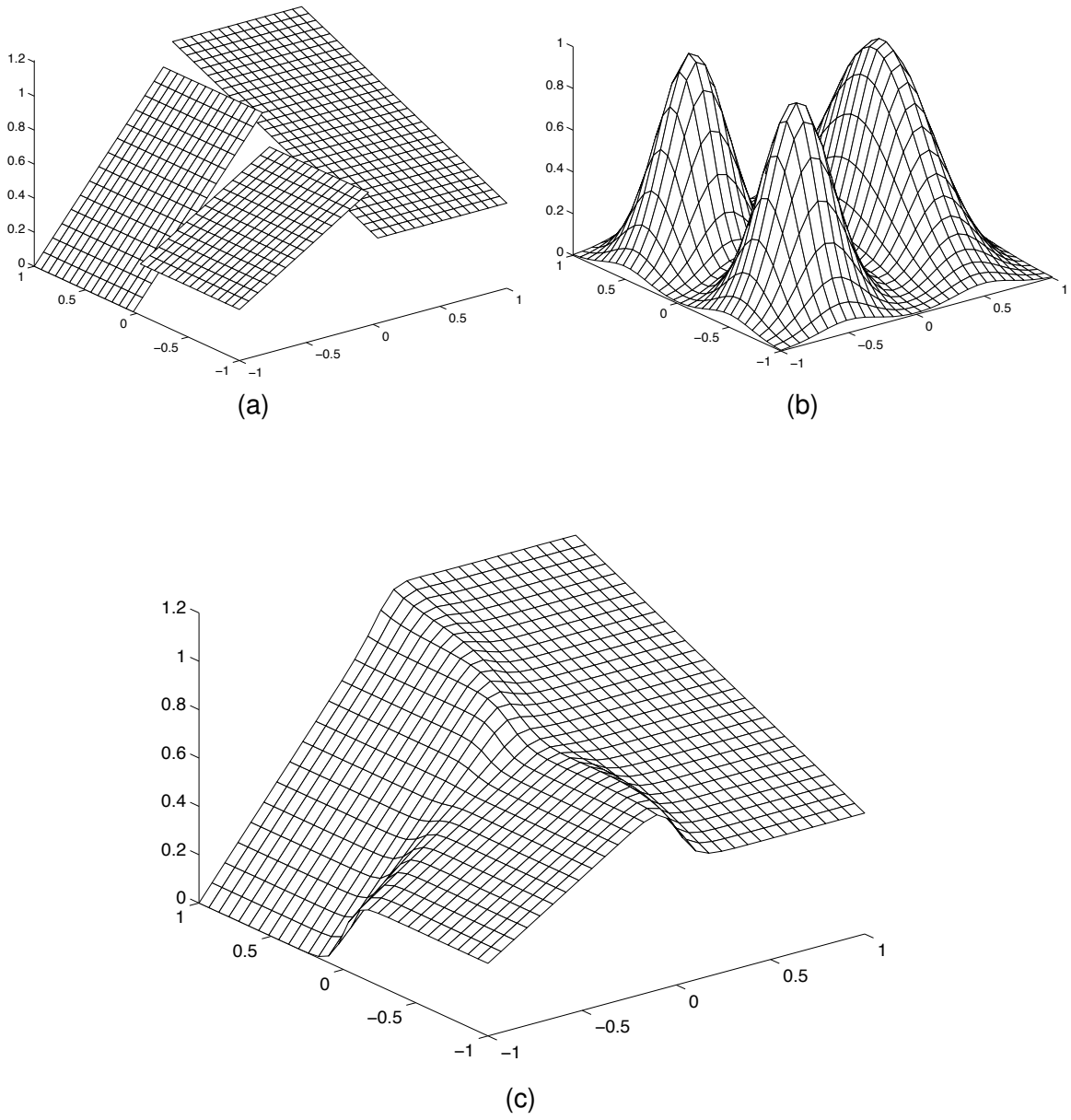


Figure 4.11: Example of Local Model Network with $m = 3$ local models. The non-linear model in (c) is obtained by combining the three local linear models in (a) according to the three basis functions in (b). Figures from [Bontempi, 1999].

composed of μ_j and Σ_j. So, $\rho_j(x)$ is the probability that, given the gene expression profile x, the breast tumor belongs to the j^{th} molecular subtype.

We can see that Equation (4.14) satisfies Equation (4.13) that is

$$\sum_{j=1}^{m} \rho_j(x) = \sum_{j=1}^{m} \frac{\pi_j \mathcal{N}(x; \mu_j, \Sigma_j)}{\sum_{l=1}^{m} \pi_l \mathcal{N}(x; \mu_l, \Sigma_l)} = 1$$

In order to reduce the complexity of the local basis functions, and so the risk of overfitting, we constraint the covariance matrices to be diagonal and equal across the Gaussians as in Equation (4.11).

It remains to estimate the parameters of the local basis functions, i.e. the number m of Gaussians, the means μ_j and the matrix of covariance Σ for the Gaussians. As described in Section 2.2.2, the parameters of the Gaussians are estimated by the EM algorithm [Dempster et al., 1977]. The number of Gaussians m is estimated using the BIC estimate [Schwarz, 1978; Fraley and Raftery, 1998].

4.3.3 Local Models

Along with the basis functions, we need to define the local risk prediction models in Equation (4.12). In this thesis, we adapt the methodology for global prognostic gene signature extraction (Section 4.1) to build the local risk prediction models. To do so, we replace the local risk prediction model $h_j(x; \alpha_j)$ in Equation (4.12) by our robust linear risk prediction model defined in Equation (4.6)

$$y = \sum_{j=1}^{m} \rho_j(x) \left[\frac{1}{k(j)} \sum_{l=1}^{k(j)} \text{sign}(\beta_l) x_l \right] \tag{4.15}$$

where x and y are the gene expression profile and the survival data of the patient respectively, $k(j)$ is the number of selected features (signature size) used to build the j^{th} local risk prediction model and β_l is the coefficient of the l^{th} feature in the corresponding univariate Cox model (see Section 4.1.3.1 for a detailed description of the risk prediction model).

We can see that the local risk prediction model in brackets in Equation (4.15) differs from the global risk prediction model in Equation (4.6) by the fact the feature ranking selecting the k most relevant features depends now on the subtype j, denoted by $k(j)$. This is made possible by an original adaptation of the feature ranking (called *local feature ranking*) to select relevant features specific to each module as presented below.

4.3.3.1 Local Feature Ranking

In order to take advantage of the modular framework, we adapt the feature ranking method presented in Section 4.1.2.1 to identify the prognostic genes in a specific module, i.e. a breast cancer molecular subtype in this thesis. Actually, we introduce a weighted version of the concordance index used as scoring function $\mathcal{S}$ in the feature ranking (Algorithm 4), in order to select genes relevant for a specific subtype. The weights were defined as ρ_j, i.e. the probability for a patient's tumor to belong to the subtype j.

Weighted Concordance Index We introduce here a weighted version of the concordance index, denoted by $C\text{-index}_{wted}$, as scoring function for the modular feature ranking. The weighted concordance index of the gene i for the subtype j takes the form

$$\mathcal{S}_{wted}(x_i, y, \rho_j) = \frac{\sum_{k,l \in \Omega} w_{kl} 1\{x_{ki} > x_{li}\}}{\sum_{k,l \in \Omega} w_{kl}} \tag{4.16}$$

where $w_{kl} = \rho_j(x_k)\rho_j(x_l)$ is the weight for the pair of patients $\{k, l\}$. So the weights are defined by the probabilities returned by the basis function ρ_j for the patients k and l. If one of the patient (or both) is unlikely to have a tumor of subtype j, the comparison of the expressions of gene i, i.e. $1\{x_{ki} > x_{li}\}$, will have a small impact on the concordance index estimation due to small weight w_{kl} in Equation (4.16). In contrast, if both patients are likely to have a tumor of subtype j, the impact of their risk scores comparison will be large.

This scoring function enables the ranking of the genes with respect to their prognostic ability in each module separately, i.e. in each breast cancer molecular subtype in this thesis. The signature size $k(j)$ for the subtype j used in Equation (4.15), is then identified through the signature stability procedure described in Section 4.1.2. The resulting local prognostic gene signatures can therefore be used to build the local risk prediction models required to compute the final risk predictions (Figure 4.9).

Pros The modular modeling approach to identify local prognostic gene signatures has several pros listed below:

- The modular modeling approach enables the use of robust linear risk prediction models to perform non-linear modeling. This is particularly important in survival analysis of microarray data where complex risk prediction models are prone to overfitting [Haibe-Kains et al., 2008c].
- From a biomedical point of view, our novel risk prediction model is easily interpretable by the doctors. Unlike global risk prediction models, this model brings two types of information relevant for the risk prediction of a patient: (i) the probabilities for a patient's tumor to belong to each breast cancer molecular subtype; and (ii) the risk predictions conditional to the subtype.
- The local prognostic gene signatures may bring more biological insights into breast cancer biology than the global ones, since these signatures take into account the molecular heterogeneity of breast cancer. In particular, different biological processes might yield good prognostic performance according to the molecular subtypes.

Cons

- A disadvantage of the modular modeling approach is that, at the same level of complexity, local models are more variant than global ones since only part of the original dataset is used for fitting the local models. In our method, some breast cancer molecular subtypes might contain only few patients being likely to have a tumor of this subtype, increasing the variance of the corresponding local models.

4.3.4 Concluding Remarks

The modular modeling approach we presented in this section, is an attempt to improve breast cancer prognostication through the development of a non-linear risk prediction model while controlling the risk of overfitting, particularly annoying in survival analysis of microarray data [Haibe-Kains et al., 2008c]. Indeed, we kept the local models robust by the use of a combination of linear univariate models (Section 4.1.3).

A key step in the modular modeling approach is the definition of the modules. This task may be challenging and may dramatically influence the performance of the final risk predictions. However, thanks to the *a priori* biological knowledge we used to develop our novel subtype clustering model, the modules can be effectively defined as the breast cancer molecular subtypes. In these settings, we will show in Section 5.4.2 that our novel risk prediction model yields good performance and significantly outperforms the state-of-the-art prognostic model in breast cancer.

4.4 A Tool for Performance Assessment and Comparison of Prognostic Gene Signatures

Performance assessment and comparison of risk prediction models are key steps in survival analysis of microarray data. A thorough performance assessment should enable the honest estimation of the quality of a model to predict the risk of new patients. Additionally, performance assessment is also extremely useful to tune the hyperparameters of methods in survival analysis (Section 4.1). A thorough performance comparison framework should allow to highlight the significant improvement of a novel method over the state-of-the-art while quantifying the increase in performance.

At the time work was begun for this thesis, only few performance criteria had been used to assess the performance of risk prediction models in breast cancer studies using microarray data (Section 3.4). Since there is no gold standard for the choice of a performance criterion, each criterion having its own advantages and disadvantages, the analyst should be able to pick up one depending on his expertise or the data to analyze. Unfortunately, the performance criteria presented in Section 2.3.5 were never implemented in the same tool to facilitate the analysis.

The field also lacks of thorough performance comparison framework (Section 3.4). Indeed, in contrast to the *naive* comparison methods, a statistical framework would enable to assess whether there is enough evidence in the data to demonstrate the superiority of a risk prediction model over another.

During the thesis, we have implemented numerous performance criteria (Sections 2.3.5 and 4.4.1) as well as a novel performance comparison framework (Section 4.4.2) in order to facilitate the performance assessment and comparison of risk prediction models. This tool has been written in R, a language and environment for statistical computing and graphics [R Development Core Team]. The R package, called `survcomp`, is fully documented and is publicly available from the comprehensive R archive network [CRAN].

In the following sections, we will present our original software contributions for the performance assessment and comparison in survival analysis.

4.4.1 Performance Assessment

Assessing the performance of a risk prediction model is not trivial since there is no gold standard for the choice of a performance criterion in survival analysis. We have seen in Section 3.4.1 that the hazard ratio is by far the most widely used performance criterion in the field of breast cancer prognostication using microarray data, despite serious drawbacks. These drawbacks raise the question of the use of alternative performance criteria in survival analysis. We propose in the `survcomp` package, an implementation of all the performance criteria described in Section 2.3.5. Compared to hazard ratio, these criteria have interesting properties:

D index: This performance criterion has two advantages compared to the hazard ratio from which the D index is derived. First, similarly to non-parametric statistics [Wasserman, 2007], it uses ranks instead of the original values of the input variable making its estimation robust to outliers. Second, its interpretation does not depend on the scale of the input variable anymore (Section 2.3.5). We refer the reader to '`?D.index`' for the details about the implementation of this performance criterion[5]. The hazard ratio from which the D index is derived is also implemented, see '`?hazard.ratio`'.

Concordance index: Unlike hazard ratio, the concordance makes no assumption about the hazard of the patients. It is robust to outliers since only the order of the risk predictions matters. Its interpretation is simple and equivalent to the area under a ROC curve, well-known in supervised classification theory [Hastie et al., 2001; Webb, 2003]. We refer the reader to '`?concordance.index`' for the details about the implementation of this performance criterion.

Time-dependent ROC curves: This performance criterion has the same advantages as the concordance index, i.e. no assumption and robust to outliers, and is particularly useful for interpretation. Indeed, the analyst is able to visually assess the trade-off between sensitivity and specificity for the classifications computed by applying each possible cutoff (Figure 4.12). The analyst is also able to assess the overall performance of an input variable by computing the area under the curve. ROC curves are widely used in supervised classification theory [Hastie et al., 2001; Webb, 2003]. We refer the reader to '`?tdrocc`' for the details about the implementation of this performance criterion.

Cross-validated partial likelihood: This performance criterion has the advantage to use the same quantity than the cost function in Cox's regression, i.e. the partial likelihood. Several authors used this criterion to tune the penalization parameter in regularized Cox's regression [Gui and Li, 2005; van Houwelingen et al., 2006]. However, there is no easy interpretation of this performance criterion and it is not robust to the presence of outliers. We refer the reader to '`?cvpl`' for the details about the implementation of this performance criterion.

Brier score: This performance criterion has a form very similar to the *mean squared error*, well-known in supervised classification theory [Hastie et al., 2001; Webb, 2003]. The main drawback of the Brier score is that it requires to estimate the baseline hazard function in case of Cox's model, which is a difficult task (Section 2.3.4.2; [Collett,

[5]Using R, one can easily access to the manual page of a function `foo` by entering the command '`?foo`'.

2003]). Moreover, this performance criterion is not robust to the presence of outliers. We refer the reader to '`?sbrier.score2proba`' for the details about the implementation of this performance criterion.

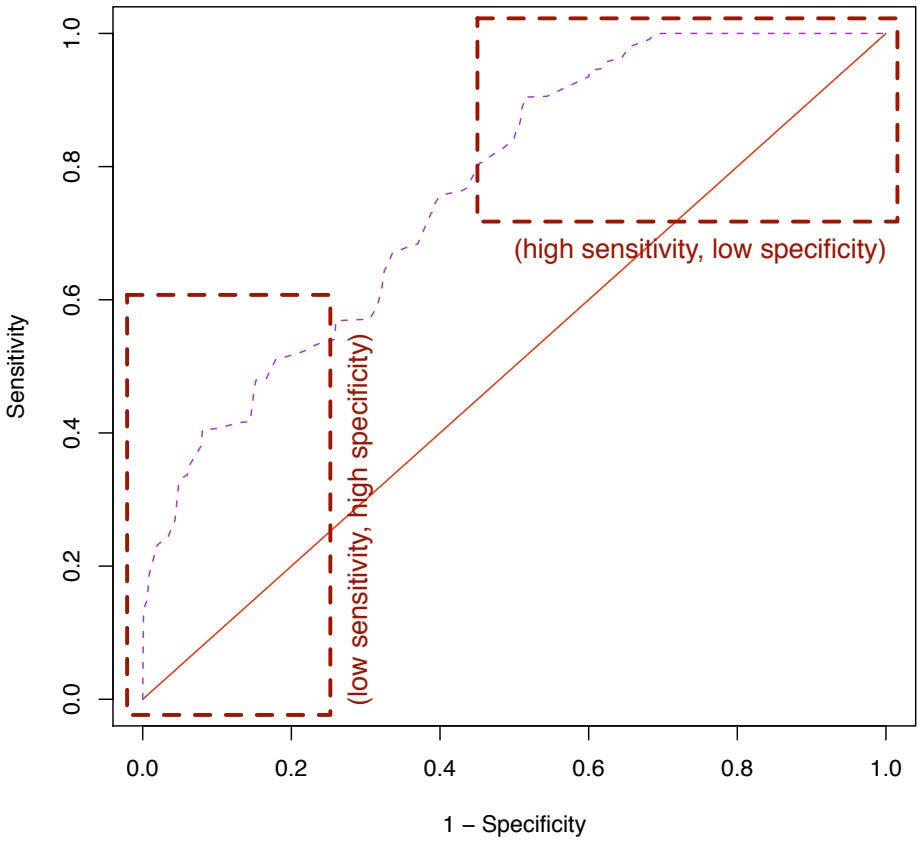


Figure 4.12: Example of ROC curves. The red diagonal line represents the performance of a random model. The violet curve represent the performance of a risk score such that large risk scores stand for high-risk patients. The two boxes illustrate the regions of the plot where different trade-offs (obtained by applying different cutoffs) can be reached.

Depending on the purpose of the risk prediction model, several criteria can be used to assess its performance and to better understand the trade-offs obtained by applying different cutoffs in case of risk group predictions from risk scores as illustrated by the time-dependent ROC curve in Figure 4.12.

4.4.2 Performance Comparison

Similarly to performance assessment, performance comparison is a key step in survival analysis of microarray data. Indeed, at the end of the comparison study, the authors should be able to claim that a new risk prediction model is competitive with the state-of-the-art or not. We have seen in Section 3.4.2 that the two main approaches to performance comparison in breast cancer prognostication studies rely on the *multivariate Cox analysis* and the *univariate*

Cox analysis and naive comparison. These two approaches have the common drawback to prevent the assessment of the significance of the superiority of a risk prediction model over another. In other words, these performance comparison methods do not answer the question whether there is enough evidence in the data to demonstrate such a superiority. This issue is particularly relevant in microarray data analysis since the sample size is often small and validation data are rarely available.

To address this issue, we propose below a novel approach allowing for statistically comparing risk predictions computed with different models.

4.4.2.1 Statistical Performance Comparison

The original framework for performance comparison we propose, borrows the principle of the *univariate Cox analysis* approach (Section 3.4.2) but makes possible the use of any performance criterion (not limited to the hazard ratio) and allows to assess the significance of a potential performance improvement.

The idea is first, to assess the performance for each risk prediction separately and second, to test the null hypothesis that the two performance estimates are equal. To do so, we used different statistical tests depending on the performance criterion. These statistical tests have been implemented in the `survcomp` package through the '*name_of_criterion*`.comp`' functions (e.g. `hazard.ratio.comp`).

Let r_1 and r_2 be the risk predictions computed by two different models from the same set of n patients. Let p_1 and p_2 be the performance estimates of r_1 and r_2 respectively.

Hazard ratio, concordance index and D index: We have seen in Section 2.3.5 that the standard error for the concordance index [Pencina and D'Agostino, 2004] and for the natural logarithm of the hazard ratio and the D index [Collett, 2003] can be computed straightforwardly. In this case, p_1 and p_2 are estimated through the concordance index or the natural logarithm of hazard ratio or D index. Let se_1 and se_2 be their respective standard errors. Assuming that p is normally distributed (see [Pencina and D'Agostino, 2004] for the concordance index and [Collett, 2003] for the hazard ratio and the D index), we can test the superiority of p_2 over p_1 using a paired Student t test

$$t_{stat} = \frac{p_1 - p_2}{\sqrt{\mathrm{se}_1^2 + \mathrm{se}_2^2 - 2\,\rho\,\mathrm{se}_1\,\mathrm{se}_2}}$$

where ρ is the correlation coefficient between the risk predictions r_1 and r_2 and t_{stat} follows a student t distribution of $n-1$ degrees of freedom. We reject the null hypothesis, that is $p_2 \leq p_1$, if the t_{stat} reaches a critical value specified by the analyst (typically a value of t_{stat} such that the corresponding one-tailed p-value is < 0.05).

Cross-validated partial likelihood, time-dependent ROC curves and Brier score: Unlike hazard ratio, concordance index and D index, the standard error of the cross-validated partial likelihood, the area under the time-dependent ROC curve and the Brier score can not be computed straightforwardly (Section 2.3.5). Therefore, we use a non-parametric test, that is the Wilcoxon signed-rank test [Wilcoxon, 1945] to test the superiority of p_2 over p_1. For the cross-validated partial likelihood, p_1 and p_2 are vectors of partial likelihoods corresponding to each fold in the cross-validation. For the area

under time-dependent ROC curve and the Brier score, p_1 and p_2 are vectors of values corresponding to each point in time t for which we observe an event occurrence. So, the Wilcoxon signed-rank test takes the form

$$T = \min\left\{\sum R^+, \sum R^-\right\}$$

where $\sum R^+$ is the sum of the ranks of the positive difference scores $p_2 - p_1 < 0$ and $\sum R^-$ is the sum of the ranks of the negative difference scores $p_2 - p_1 \leq 0$. The T statistic is interpreted by employing a table of critical T values reported in [Wilcoxon, 1945].

Additionally to demonstrate the superiority of a risk prediction model over another, the analyst is usually interested in highlighting the potential complementarity of these two models, two models being complementary if their combination in a new prediction model yields better performance than the two models separately. In our performance comparison framework, the complementarity between two risk predictions can be shown similarly to the *univariate Cox analysis* approach for performance comparison (Section 3.4.2). Indeed, the analyst can build a model combining the risk predictions and assess the performance of such a model. If the performance of this model is significantly better than for each risk prediction separately, this suggests that the risk predictions are complementary.

4.4.3 Report for Large Comparative Studies

In case of a large comparative study with numerous risk prediction models, datasets and performance criteria, it may be difficult to report comprehensively a large amount of results. To do so, we use a textual and graphical representation adapted from the literature.

4.4.3.1 Textual Representation

The textual representation of the results simply reports the estimates of the performance criteria in a table such that a performance significantly better than a benchmark, i.e. a risk prediction model from the state-of-the-art, is written in **bold** face. This allows for easily pointing out the risk prediction models which outperform consistently the benchmark with respect to the performance criterion and the dataset under study. Such a textual representation of the results from the performance assessment and comparison is illustrated in Table 4.1. In this example, M1 outperforms the benchmark model only in the dataset D2 for the D index and the IAUC. On the contrary, M2 is always significantly better than the benchmark model in the dataset D2 and only for the concordance index in the dataset D1.

4.4.3.2 Graphical Representation

The standard error of the concordance index and of the natural logarithm of the hazard ratio and D index can be computed straightforwardly (Section 2.3.5). Since we assumed that these estimates follow a normal distribution, we can compute their $(100 - \alpha)$% confidence interval as $p \pm z_{\alpha/2}\text{se}(p)$, where p is the performance estimate and $z_{\alpha/2}$ is the upper $\alpha/2$-point of the standard normal distribution.

Model	*C*-index		D index		IAUC		IBSC	
	D1	D2	D1	D2	D1	D2	D1	D2
Benchmark	0.636	0.609	2	1.92	0.683	0.601	0.178	0.144
M1	0.606	0.618	2.13	**2.4**	0.602	**0.643**	0.185	0.143
M2	**0.658**	**0.633**	2.33	**3.02**	0.688	**0.637**	0.164	**0.131**

Table 4.1: Example of textual representation of results from performance assessment and comparison of three risk prediction models (Benchmark, M1 and M2) in two independent datasets (D1 and D2) using four performance criteria (*C*-index, D index, IAUC and IBSC, see Section 2.3.5).

To graphically represent a large number of performance estimates and their confidence interval, we use a forestplot [Lewis and Clarke, 2001] as illustrated in Figure 4.13. The performances are shown as squares centered on the point estimate of the performance of each model. A horizontal line runs through the square to show its 95% confidence interval. The vertical grey line represents the performance of a null model. The red vertical line is centered on the point estimate of the performance of the benchmark model. The p-value on the right side of each performance is the significance of the superiority of the corresponding model over the benchmark model.

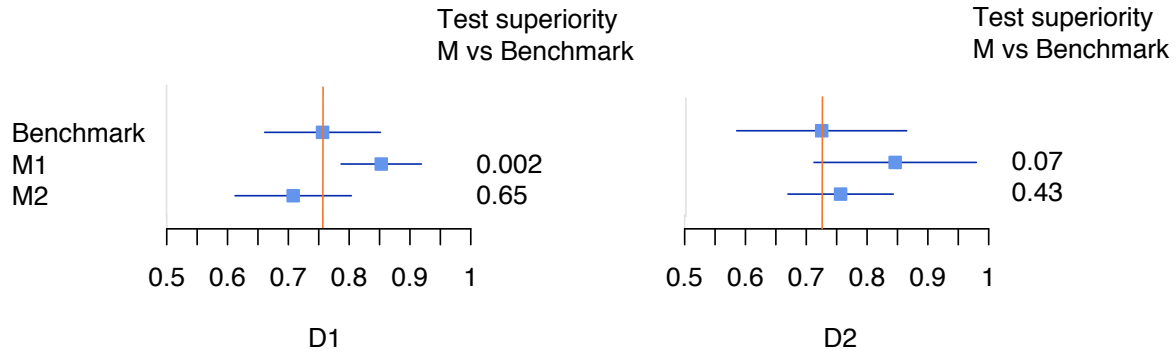


Figure 4.13: Example of a graphical representation of the concordance index performance criterion using a forestplot. The vertical grey line represents the performance of a null model. The red vertical line is centered on the point estimate of the performance of the benchmark model. The p-value on the right side of each performance is the significance of the superiority of the corresponding model over the benchmark model. In this case, M1 significantly outperforms the benchmark model in dataset D1 and tend to be better in dataset D2 (but it does not reach statistical significance, p-value ≥ 0.05). In contrast, M2 is not significantly better than the benchmark model whatever the dataset.

For the IAUC and the IBSC performance criteria (Section 2.3.5), we can easily represent the evolution of the area under the curve (AUC) and the Brier score (BSC) respectively, with respect to the time as illustrated in Figures 4.14 and 4.15.

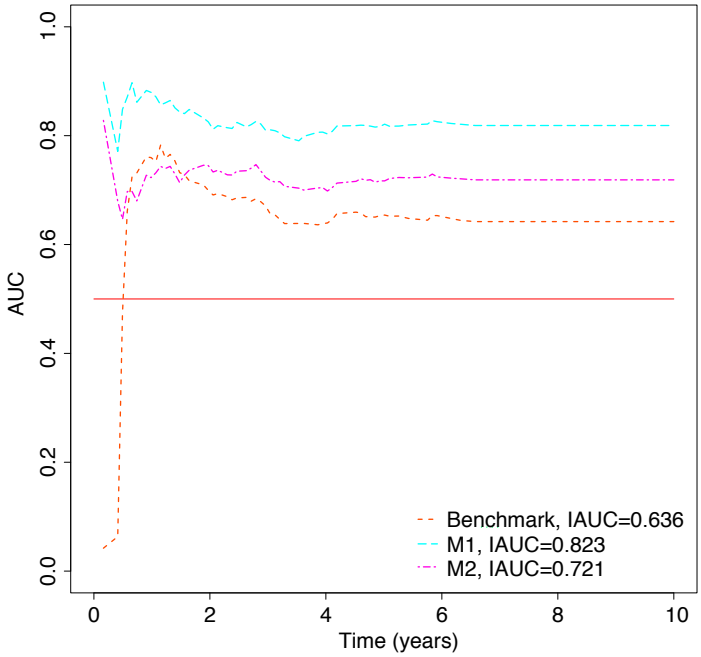


Figure 4.14: Example of a graphical representation of the IAUC performance criterion.

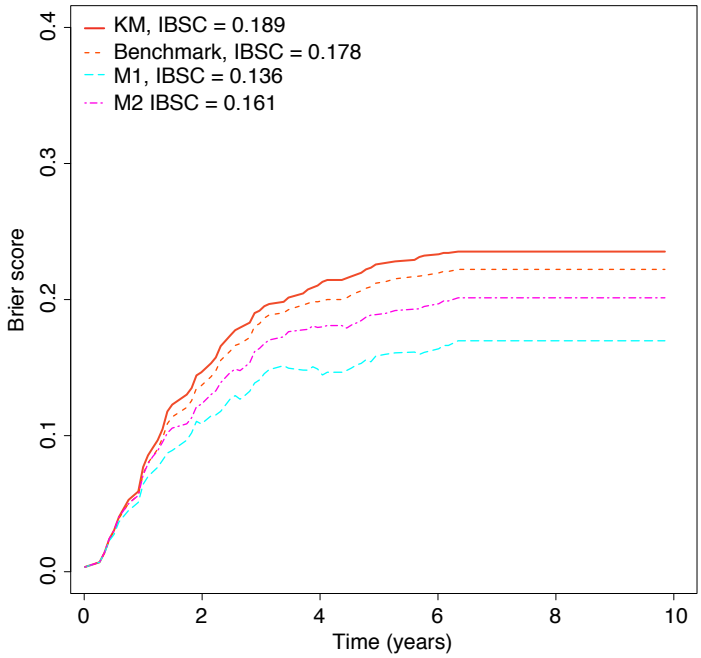


Figure 4.15: Example of a graphical representation of the IBSC performance criterion.

4.4.4 Concluding Remarks

We proposed in this section an original framework for performance assessment and comparison of risk prediction models. This framework was implemented in an R package, called `survcomp`. This implementation allows the analyst to choose one or more criteria to assess the performance of risk prediction models using the same tool, thus facilitating the analysis. In addition, we proposed a textual and a graphical representations of the results obtained from large comparative studies as in [Haibe-Kains et al., 2008c].

Chapter 5

Experimental Findings

This chapter details the experimental findings which were made possible by means of the original methods presented in this thesis. The outline of the chapter is as follows: first we introduce in Section 5.1 the numerous microarray datasets we collected in order to develop prognostic gene signatures. Then, in Section 5.2 we report the experimental findings resulting from the development of two global prognostic gene signatures. In spite of the good performance of these signatures, we sought to improve them by integrating the biological knowledge related to the breast cancer molecular subtypes into their development. We report in Section 5.3 the experimental findings related to our novel subtype clustering model. Finally, using this clustering model, we report in Section 5.4 the experimental findings derived from the development of several local prognostic gene signatures.

For each specific study, the experimental findings are presented according to the following structure:

1. Motivations of the study.
2. Methods and hyperparameters used in the analysis.
3. Results:
   (a) The gene signature extracted using the methods presented previously, description of the data, and the intermediate results obtained during signature extraction.
   (b) Performance assessment of the gene signature using the performance criteria discussed in Section 4.4.1.
   (c) Performance comparison between the gene signature and the state-of-the-art, based on the statistical framework presented in Section 4.4.2.
4. Findings of the study.

In the next sections, the main experimental findings are summarized, and the reader is directed to additional results published in the original articles.

5.1 Datasets

Early breast cancer microarray studies lacked data for extracting a gene signature and validating it. This scarcity of data is due to the high cost of microarray technology and the

scarcity of frozen tumor tissues. To address this problem, we collected, with the help of Dr. Pratyaksha Wirapati from the Bioinformatics Core Facility (Swiss Institute for Experimental Cancer Research, Lausanne, Switzerland), numerous breast cancer microarray datasets. These datasets were retrieved from

- Original author's website.
- Journal article's supplementary materials.
- Public repositories:
  - Gene Expression Omnibus (GEO; [Barrett et al., 2005]).
  - ArrayExpress (AE; [Parkinson et al., 2005]).
  - Stanford Microarray Database (SMD; [Hubble et al., 2009]).
- Third party curators:
  - Cleanex [Praz et al., 2003].
  - Oncomine [Rhodes et al., 2007].

Data collection and preparation is a tedious task involving numerous steps before it is possible to effectively analyze the data. First, a comprehensive survey has to be conducted to find new datasets related to the biomedical question of interest. This survey may be complicated by the fact that parts of the same dataset may be in different places, e.g. author/journal's website for the clinical information and public repositories for microarray data. Occasionally, some parts of a dataset may be unavailable, especially clinical information. Therefore, one often has to request the help of the author(s) to retrieve all the necessary information. Once all the data are available, tedious manual clean-up and reformatting is often required. The annotations of the microarray platforms need to be updated, since knowledge about the transcriptome is still evolving. Finally, after all the data are collected, the microarray data and the corresponding clinical information are stored using the same format to facilitate large scale analysis.

For each dataset, several types of survival data are reported depending on the type of event being recorded in the clinical study:

- The relapse free survival (RFS) refers to survival data for which the event is either the appearance of a relapse or the death of a patient from any cause.
- The distant metastasis free survival (DMFS) refers to survival data for which the event is either the appearance of a distant metastasis or the death of a patient from any cause.
- The overall survival (OS) refers to survival data for which the event is the death of a patient from any cause.

In this thesis, we focus on DMFS (or RFS in case of missing DMFS information), since these survival data are the most adequate to study the relationship between the gene expression profile of a primary tumor and its evolution or response to treatment.

All the datasets we collected are briefly described below and summarized in Table 5.1.

NKI: This dataset was used for extracting the GENE70 signature in the first study conducted for breast cancer prognostication using microarray technology [van't Veer et al., 2002] followed by its validation [van de Vijver et al., 2002]. The microarray technology used in this study is Agilent with a chip of 24,481 probes representing 13,120 unique genes for 345 heterogeneously treated patients. The tissue samples are from the Nederlands Kanker Instituut (Amsterdam, The Netherlands). RFS, DMFS and OS data are available.

STNO2: This dataset recapitulates the microarray data used in the first studies of breast cancer molecular profiling [Sorlie et al., 2003]. The microarray technology used in this study is a cDNA microarray developed at Stanford with a chip of 7,787 probes representing 5,427 unique genes for 122 heterogeneously treated patients. The tissue samples are from the Stanford University (Stanford, USA) and the Norwegian Radium Hospital (Oslo, Norway). RFS and OS data are available.

NCI: This dataset was used in another seminal work confirming the presence of molecular subtypes in breast cancer in a population-based study [Sotiriou et al., 2003]. The microarray technology used in this study is a cDNA microarray developed at the National Cancer Institute (Bethesda, USA) with a chip of 6,878 probes representing 4,112 unique genes for 99 heterogeneously treated patients. The tissue samples are from the John Radcliffe Hospital (Oxford, United Kingdom). RFS data are available.

MGH: This dataset was used to extract a gene signature predictive of the resistance to tamoxifen [Ma et al., 2004]. The microarray technology used in this study is Arcturus with a chip of 22,575 probes representing 11,421 unique genes for 60 ER-positive patients homogeneously treated by tamoxifen. The tissue samples are from Massachusetts General Hospital (Boston, USA). RFS and DMFS data are available.

MSK: This dataset was used to study the genes that mediate breast cancer metastasis to the lung [Minn et al., 2005]. The microarray technology used in this study is Affymetrix with a chip of 22,283 probes representing 11,837 unique genes for 99 heterogeneously treated patients. The tissue samples are from the Memorial Sloan-Kettering Cancer Center (New York, USA). DMFS data are available.

UPP: This dataset was used to extract the P53 signature for the P53 mutations status in breast cancer [Miller et al., 2005] . The microarray technology used in this study is Affymetrix with a chip of 22,283 probes representing 11,837 unique genes for 251 heterogeneously treated patients. The tissue samples are from the Karolinska Institute (Uppsala, Sweden). RFS data are available.

STK: This dataset was used to extract a prognostic gene signature [Pawitan et al., 2005]. The microarray technology used in this study is Affymetrix with a chip of 22,283 probes representing 11,837 unique genes for 159 heterogeneously treated patients. The tissue samples are from the Karolinska Institute (Stockholm, Sweden). RFS data are available.

VDX: This dataset was used to extract the GENE76 signature in the second large study for prognostication of early breast cancer [Wang et al., 2005]. A small set of ER-negative patients were added to this dataset to study the metastatic spread of breast tumors

into lung [Minn et al., 2007]. The microarray technology used in this study is Affymetrix with a chip of 22,283 probes representing 11,837 unique genes for 344 node-negative untreated patients. The tissue samples are from Erasmus Medical Center (Rotterdam, The Netherlands). RFS and DMFS data are available.

UNT: This dataset was used to understand the molecular basis of histological grade and for extracting the gene expression grade index signature [Sotiriou et al., 2006]. The microarray technology used in this study is Affymetrix with a chip of 22,283 probes representing 11,837 unique genes for 137 node-negative untreated patients. The tissue samples are from the John Radcliffe Hospital (Oxford, United Kingdom) and the Karolinska Institute (Uppsala, Sweden). RFS and DMFS data are available.

UNC2: This dataset was used to study EGFR expression profile according to the breast cancer molecular subtypes [Hoadley et al., 2007]. The microarray technology used in this study is Agilent with a chip of 21,495 probes representing 10,157 unique genes for 248 heterogeneously treated patients. The tissue samples are from the University of North Carolina at Chapel Hill (Chapel Hill, USA). RFS and OS data are available.

DUKE: This dataset was used to study the prognostic value of oncogenic pathway signatures extracted from cell line experiments [Bild et al., 2006]. The microarray technology used in this study is Agilent with a chip of 12,625 probes representing 8,149 unique genes for 171 heterogeneously treated patients. The tissue samples are from the Duke University (Durham, USA). OS data are available.

CAL: This dataset was used to study genomic and transcriptional aberrations linked to breast cancer pathophysiologies [Chin et al., 2006]. The microarray technology used in this study is Affymetrix with a chip of 22,283 probes representing 11,837 unique genes for 118 heterogeneously treated patients. The samples are from the University of California and from the California Pacific Medical Center (San Francisco, USA). RFS, DMFS and OS data are available.

TBG: This dataset was used to validate the GENE76 and GENE70 signatures [Desmedt et al., 2007]. The microarray technology used in this study is Affymetrix with a chip of 22,283 probes representing 11,837 unique genes for 198 node-negative untreated patients. The tissue samples are from the John Radcliffe Hospital (Oxford, United Kingdom), Guy's Hospital (London, United Kingdom), the Karolinska Institute (Uppsala, Sweden), the René Huguenin Hospital (Saint-Cloud, France) and the Gustave Roussy Institute (Villejuif, France). RFS, DMFS and OS data are available.

NCH: This dataset was used to extract a gene signature for breast cancer prognostication [Naderi et al., 2007]. The microarray technology used in this study is Agilent with a chip of 17,086 probes representing 13,784 unique genes for 135 heterogeneously treated patients. The tissue samples are from Nottingham City Hospital (Nottingham, United Kingdom). RFS, DMFS and OS data are available.

DUKE2: This dataset was used to validate gene signatures predictive of response to neoadjuvant chemotherapy [Bonnefoi et al., 2007]. The microarray technology used in this study is Affymetrix with a chip of 61,359 probes representing 16,559 unique genes for 160 patients homogeneously treated by chemotherapy. The tissue samples are from Duke University (Durham, USA). No survival data are available.

MAINZ: This dataset was used to study the prognostic value of the humoral immune system in breast cancer [Schmidt et al., 2008]. The microarray technology used in this study is Affymetrix with a chip of 22,283 probes representing 11,837 unique genes for 200 node-negative untreated patients. The tissue samples are from the Johannes Gutenberg University Mainz (Mainz, Germany). DMFS data are available.

TAM: This dataset was used to extract a gene signature predictive of the resistance to tamoxifen [Loi et al., 2007]. The microarray technology used in this study is Affymetrix with a chip of 44,928 probes representing 15,684 unique genes for 354 ER-positive patients homogeneously treated by tamoxifen. The tissue samples are from the John Radcliffe Hospital (Oxford, United Kingdom), the Karolinska Institute (Uppsala, Sweden) and Guy's Hospital (London, United Kingdom). RFS and DMFS data are available.

TAM2: This dataset was used to extract a gene signature predictive of resistance to tamoxifen [Chanrion et al., 2008]. The microarray technology used in this study is Aminolink with a chip of 21,332 probes representing 14,031 unique genes for 155 ER-positive patients homogeneously treated by tamoxifen. The tissue samples are from the Cancer Research Center of Val d'Aurelle (Montpellier, France), the Bergonie Institute (Bordeaux, France) and the Department of Obstetrics and Gynecology of Turin (Turin, Italy). RFS, DMFS and OS data are available.

LUND2: This dataset was used to study the prognostic value of a gene signature for PTEN tumor suppressor pathway activity [Saal et al., 2007]. The microarray technology used in this study is Swegene with a chip of 27,648 probes representing 12,288 unique genes for 105 patients homogeneously treated by tamoxifen. The tissue samples are from the Lund University Hospital (Lund, Sweden). No survival data are available.

LUND: This dataset was used to extract a gene signature predictive of resistance to radiotherapy in breast cancer [Nimeus-Malmstrom et al., 2008]. The microarray technology used in this study is Swegene with a chip of 26,824 probes representing 14,676 unique genes for 143 heterogeneously treated patients. The tissue samples are from the Lund University Hospital (Lund, Sweden). No survival data are available.

MUG: This dataset was used to study the effects of infiltrating lymphocytes and estrogen receptors on gene expression and prognosis in breast cancer [Calabrò et al., 2008]. The microarray technology used in this study is Operon with a chip of 35,788 probes representing 16,783 unique genes for 152 heterogeneously treated patients. The tissue samples are from the Medical University of Graz (Graz, Austria). No survival data are available.

In summary, all the datasets are composed of gene expressions from primary breast tumor tissues. Most of them are representative of a global population of breast cancer patients, except for MGH, TAM and TAM2, in which only ER-positive patients were selected. Treatment information and survival data are not available for some datasets (e.g. DUKE2, LUND2, LUND or MUG).

The performance assessment and comparison (Sections 3.4 and 4.4) of prognostic gene signatures are key steps in microarray studies dealing with breast cancer prognostication. It is worth to note that, in an ideal situation, the performance assessment and comparison

Article	Dataset	Technology	Survival	Treatment	Patients	Probes
[van't Veer et al., 2002; van de Vijver et al., 2002]	NKI	Agilent	RFS, DMFS, OS	untreated, chemo	345	24,481
[Sorlie et al., 2003]	STNO2	cDNA Stanford	RFS, OS	untreated, chemo, hormono	122	7,787
[Sotiriou et al., 2003]	NCI	cDNA NCI	RFS	untreated, chemo, hormono	99	6,878
[Ma et al., 2004]	MGH	Arcturus	RFS, DMFS	hormono	60	11,421
[Minn et al., 2005]	MSK	Affymetrix	DMFS	heterogeneous	99	22,283
[Miller et al., 2005]	UPP	Affymetrix	RFS	untreated, hormono	251	22,283
[Pawitan et al., 2005]	STK	Affymetrix	RFS	untreated, chemo, hormono	159	22,283
[Wang et al., 2005; Minn et al., 2007]	VDX	Affymetrix	RFS, DMFS	untreated	344	22,283
[Sotiriou et al., 2006]	UNT	Affymetrix	RFS, DMFS	untreated	137	22,283
[Hoadley et al., 2007]	UNC2	Agilent	RFS, OS	heterogeneous	248	21,495
[Bild et al., 2006]	DUKE	Affymetrix	OS	heterogeneous	171	12,625
[Chin et al., 2006]	CAL	Affymetrix	RFS, DMFS, OS	chemo, hormono	118	22,283
[Desmedt et al., 2007]	TBG	Affymetrix	RFS, DMFS, OS	untreated	198	22,283
[Naderi et al., 2007]	NCH	Agilent	RFS, DMFS, OS	untreated, chemo, hormono	135	17,086
[Bonnefoi et al., 2007]	DUKE2	Affymetrix	NA	chemo	160	61,359
[Schmidt et al., 2008]	MAINZ	Affymetrix	DMFS	untreated	200	22,283
[Loi et al., 2007]	TAM	Affymetrix	RFS, DMFS	hormono	354	44,928
[Chanrion et al., 2008]	TAM2	Aminolink	RFS, DMFS, OS	hormono	155	21,332
[Saal et al., 2007]	LUND2	Swegene	NA	hormono	105	27,648
[Nimeus-Malmstrom et al., 2008]	LUND	Swegene	NA	heterogeneous	143	26,824
[Calabrò et al., 2008]	MUG	Operon	NA	NA	152	16,783

Table 5.1: Table describing all the datasets of patients used in the experiments presented in this thesis. Legend: RFS = Relapse Free Survival: DMFS = Distant Metastasis Free Survival; OS = Overall Survival; untreated = no treatment; chemo = chemotherapy; hormono = hormonotherapy; heterogeneous = heterogeneous treatments: NA = Not Available.

should be performed on a large independent dataset. If the analyst has to develop the risk prediction model and to validate it on the same dataset, a cross-validation framework could be used [Stone, 1974]. If independent datasets are readily available, the analyst should use as many datasets as possible in order to increase the sample size for validation.

Thanks to the large number of datasets collected in this thesis, we always used independent datasets to assess and to compare the performance of our novel prognostic gene signatures presented in the next sections.

5.2 Global Prognostic Gene Signatures

Using the methodology described in Section 4.1, we developed two global prognostic gene signatures, namely, the gene expression grade index (GGI) and the tamoxifen resistance signature (TAMR13).

5.2.1 Gene Expression Grade Index (GGI)

5.2.1.1 Motivations

The relationship between traditional histo-pathological parameters and gene expressions was barely known in early 2000s. Since histo-pathological parameters were previously shown to be highly prognostic, it would be interesting to study their molecular basis with the hope to improve their measurement and so their prognostic value. In this section, we will present the experimental findings of a study of the molecular basis of histological grade.

Histological grade, described in Section 1.2, is one the most prognostic clinical variables in breast cancer, discriminating patients at low, intermediate and high-risk of recurrence as histological grade 1, 2 and 3 respectively. Using gene expression data, we sought to identify the genes that are differentially expressed between histological grade 1 and 3 tumors. The aim of this study was:

- To better understand the molecular basis of histological grade.
- To reclassify intermediate histological grade tumors (histological grade 2) into histological grade 1 or 3-like tumors.
- To improve the quantification of tumor differentiation in order to yield better breast cancer prognostication.

The experimental findings of this study, which we present below, were published in [Sotiriou et al., 2006].

5.2.1.2 Methods

In order to extract a gene signature predictive of histological grade, we used the methodology presented in Section 4.1:

1. Feature transformation: The original gene expressions were used for the following steps of signature extraction, with no feature transformation being performed.

2. Feature selection: The gene expressions were ranked with respect to their differential expression between histological grade 1 and 3. Since the prediction outcome is a binary class coding for histological grade 1 and 3, we used the significance of the standardized mean difference [Hedges and Olkin, 1987] as the scoring function to rank the genes. The p-values were corrected for multiple testing using the maxT algorithm adapted for high dimensional data [Korn et al., 2004]. The number of genes in the signature was selected by using a threshold p-value of 0.05 for a false discovery count of 0.

3. Model building: Once the signature was identified, we built a predictive model (Section 4.1.3.1), which was a signed average of the expressions of the genes included in the signature. The resulting score was called the gene expression grade index (GGI). We scaled the GGI values using histological grade such that the mean of the GGI values for histological grade 1 and 3 tumors equaled to -1 and +1 respectively. This ensured the GGI values having approximately the same scale for different populations of breast cancer patients and for different microarray technologies. We referred to GG as the dichotomized version of the GGI using a cutoff of 0, the middle point between mean GGI values of histological grade 1 and 3 tumors.

5.2.1.3 Results

Gene signature As a training set, we used 64 ER-positive tumors from tamoxifen treated patients (subset of TAM dataset, see Table 5.1). These tumors had either a histological grade 1 or 3. We used only ER-positive tumors for selecting the genes because of the dependence between ER status and histologic grade: almost all ER-negative tumors are classified as either histologic grade 2 or 3. If we had used all histologic grade 1 and 3 tumors regardless of the ER status in our training set, we would have selected ER-related genes that were spuriously associated with grade. The fact that these tumors came from patients being treated by tamoxifen is not an issue since we used only information on ER status and histologic grade to extract the gene signature, without considering the clinical outcome. For microarray profiling and grading, we used primary tumor tissues that were collected before the beginning of tamoxifen treatment; consequently, the gene signature identified with the training set is not affected by disease outcome or treatment.

The signature extracted from this training set is composed of 128 probes representing 97 unique genes (Appendix C.1). The expression pattern of this signature was fairly homogenous in the training set (see Figure 1.A in [Sotiriou et al., 2006]). Most genes are over-expressed in histological grade 3 tumors and have biological functions that have been previously associated with cell cycle progression and proliferation (among the top 20 over-expressed genes are UBE2C, KPNA2, TPX2, FOXM1, STK6, CCNA2, BIRC5, and MYBL2, well-known in the literature).

In order to test whether the histological grade 2 tumors have a gene expression profile that is distinctive from histological grade 1 and 3 tumors, we attempted to extract a signature as we did previously. Interestingly, we found that no gene passed the threshold (p-value $<$ 0.05 for a false discovery count of 0), suggesting that either the gene expression profile of histological grade 2 tumors is intermediate between those from histological grade 1 and 3 tumors or is a heterogeneous mixture of profiles from histological grade 1 and 3 tumors. In fact, we can see in the heatmaps of the signature in some independent datasets

(see Figures 1.B-1.E in [Sotiriou et al., 2006]) that the vast majority of histological grade 2 tumors exhibits a gene expression profile very similar to either histological grade 1 or 3 tumors, whereas only a small proportion with GGI close to zero, exhibit an intermediate gene expression profile.

Performance assessment We assessed performance from a classification (histological grade) and a prognostic (survival) point of view. Given that several microarray datasets have been made publicly available since the publication of this study, we decided to update the results published in [Sotiriou et al., 2006] by using new independent datasets. In this section, we used the set of untreated node-negative breast cancer patients included in the NKI, TBG, UPP, UNT and MAINZ datasets. This avoided introducing the confounding factor of treatment heterogeneity, which may lead to misleading results from a prognostic point of view.

Histological grade We validated the performance of the GGI to predict histological grade (1 vs 3) by computing the ROC curve and its corresponding AUC in the independent set of untreated node-negative breast cancer patients (Figure 5.1). The performance was excellent, the ROC curve of the GGI having an AUC of 0.88. This AUC is significantly better than the AUC of the null model represented by the diagonal (p-value $<$ 1E-16).

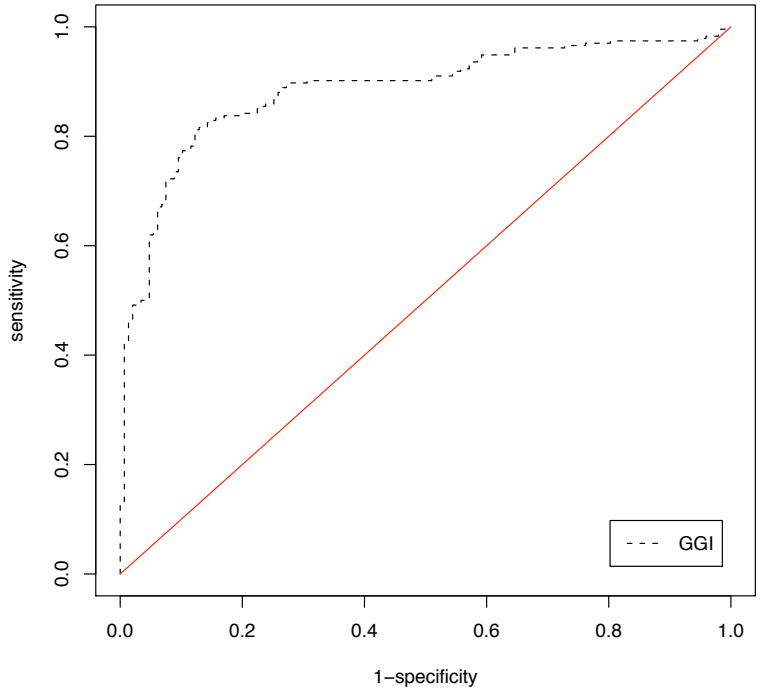


Figure 5.1: ROC curve of the GGI predicting the histological grade (1 or 3) of patients in the independent dataset of untreated node-negative patients (NKI, TBG, UPP, UNT and MAINZ).

Prognosis Using DMFS as survival endpoint, we reproduced the Figures 2.A-C published in [Sotiriou et al., 2006] but using the independent dataset of untreated node-negative

breast cancer patients. Remarkably, the updated results, sketched in Figure 5.2, were virtually identical to those in the paper. The only notable difference lies in the fact that the survival of the untreated node-negative patients was better than that reflected in the published survival curves. This was due to the use, in the paper, of datasets that included a large proportion of high-risk treated patients, for example, in STNO (subset of STNO2, see table 5.1). Figure 5.2 (a) illustrates the good prognostic ability of histological grade, as expected from previous reports [Scarff and Torloni, 1968], as well as the large proportion of histological grade 2 [Elston and Ellis, 1991]. Figure 5.2 (b) shows that GG, the dichotomized version of the GGI, was able to discriminate patients with histological grade 2 tumors into low-risk (GG1) and high-risk (GG3) groups that exhibited similar survival to that of patients with histological grade 1 and 3 tumors respectively. This result suggests that the GGI is able to re-classify histological grade 2 tumors into groups similar to histological grade 1 and 3. Lastly, Figure 5.2 (c) illustrates the prognostic ability of the GG in discriminating low-risk and high-risk groups of patients while avoiding to classify patients into an intermediate-risk group, the latter being particularly annoying in clinical practice.

5.2.1.4 Findings

The main findings of this study are twofold:

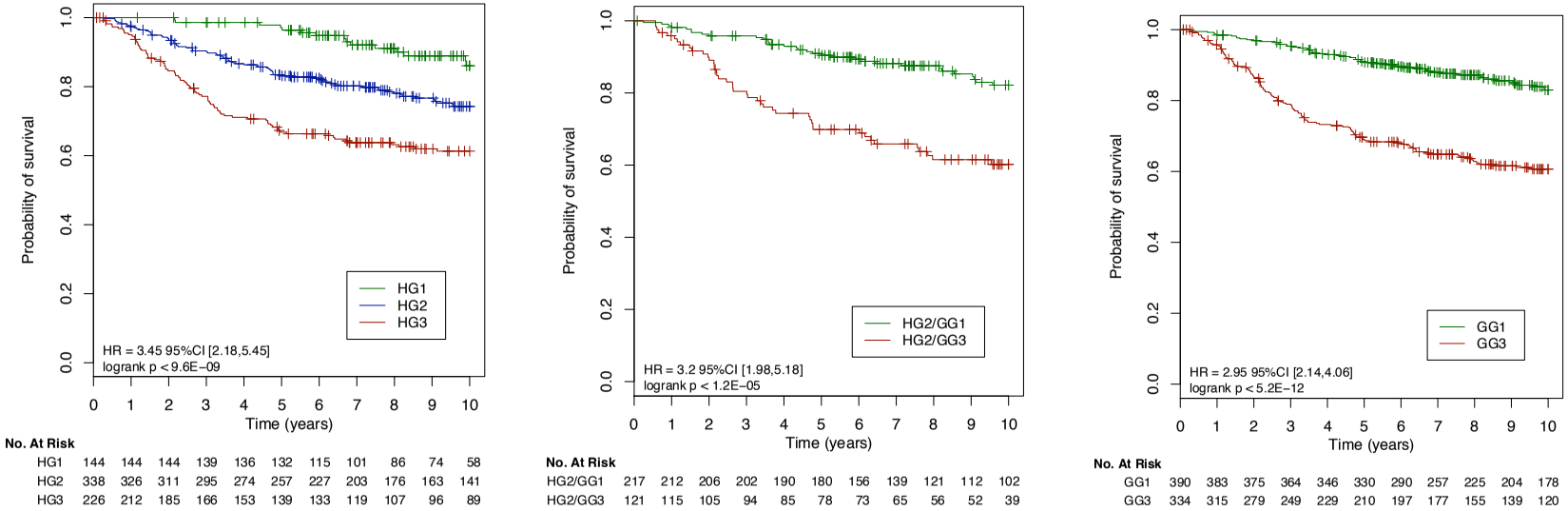
- Histological grade: We found that histologic grades 1 and 3 breast cancers have distinct gene expression profiles, but that histologic grade 2 tumors have heterogeneous gene expression profiles that range from those for histologic grade 1 tumors to those for histologic 3 grade tumors. Only a small proportion of histological grade 2 tumors exhibits an intermediate profile with GGI values close to 0.
- Prognosis: We investigated the clinical implications of the previous findings and discovered that, like the histological grade, the GGI is strongly associated with distant metastasis free survival. Our most important observation was that the three-category histologic grading system could be replaced with a two-category one that may be more clinically relevant. Thus, this grading system has the potential to improve the accuracy of grading for prognostic purposes [Sotiriou et al., 2006].

In the future, a minimal set of genes from the GGI signature should be defined that can accurately divide histologic grade 2 tumors into prognostically distinct groups. Because the genes expressed are highly correlated with one another, arbitrary subsets of the signature that are chosen only by technical constraints (such as the abundance of the RNA transcripts or the signal strength from specific probes) might be chosen to develop a practical diagnostic system using a cheaper gene expression profiling technology, such as the reverse transcription polymerase chain reaction (RT-PCR). Preliminary results in this direction have recently been published [Durbecq et al., 2007; Toussaint et al., 2008].

5.2.2 Performance Comparison of the Gene Expression Grade Index (GGI)

5.2.2.1 Motivations

An important issue in the identification of prognostic gene signatures is the comparison with state-of-the-art methods, existing gene signatures and traditional prognostic models (Section 1.2.1). A comparative study involving the GGI is particularly interesting:



	3 yrs	5 yrs	10 yrs
HG1	0.99	0.96	0.86
HG2	0.90	0.83	0.74
HG3	0.76	0.66	0.61

(a)

	3 yrs	5 yrs	10 yrs
HG2/GG1	0.95	0.90	0.82
HG2/GG3	0.80	0.70	0.60

(b)

	3 yrs	5 yrs	10 yrs
GG1	0.95	0.91	0.83
GG3	0.79	0.69	0.61

(c)

Figure 5.2: Survival analysis of histological grade and the GGI in the independent datasets of untreated node-negative breast cancer patients (NKI, TBG, UPP, UNT and MAINZ): (a) Survival curves stratified by histological grade (HG); (b) Survival curves of patients with histological grade 2 tumors, stratified by gene expression grade (GG); (c) Survival curves stratified by gene expression grade (GG). The difference in survival between groups is summarized by the hazard ratio (HR), and its significance is estimated by the logrank test. The tables report the probability of survival for each strata at three, five and ten years.

- From a statistical point of view: Is the simple prognostic model used for the GGI competitive with more complex prognostic models? Can we improve breast cancer prognostication by fitting multivariate and/or non-linear prediction models?
- From a biological point of view: Since the GGI signature is primarily composed of genes known to be related to cell proliferation, are genes related to other biological processes needed to yield better breast cancer prognostication?

Such a comparative study has been the subject of two publications [Haibe-Kains et al., 2008b,c]. In the study described below, we showed that (i) the simple risk prediction model of GGI yields excellent performance compared to more complex methods for breast cancer prognostication [Haibe-Kains et al., 2008c]; and (ii) the GGI signature is competitive with state-of-the-art prognostic signatures (namely GENE70 and GENE76) [Haibe-Kains et al., 2008b], highlighting the importance of proliferation in breast cancer prognostication.

5.2.2.2 Methods

First, we performed a large-scale comparative study of risk prediction models to compare simple to complex models with a single proliferation gene (AURKA). We considered 13 risk prediction models, including the GGI, as described in Table 1 of [Haibe-Kains et al., 2008c]. In order to elucidate the key characteristics of successful risk prediction models, we used our novel performance assessment and comparison framework (Section 4.4) with VDX as the training set and TBG, TAM and UPP as independent datasets (Table 5.1).

Second, we compared the performance of the GGI signature with state-of-the-art gene signatures, namely GENE70 and GENE76. Since our laboratory was involved in a TRANSBIG study of which the aim was to validate the GENE70 and the GENE76 signatures using the original microarray platforms and algorithms [Buyse et al., 2006; Desmedt et al., 2007], we had a unique opportunity to objectively compare these two signatures and the GGI. Indeed, the TBG dataset is the only dataset for which the original microarray platforms, i.e. Agilent for GENE70 and Affymetrix for GENE76, and the original algorithms (the risk group predictions were computed by the investigators of the signatures) were used to assess the performance of the two signatures. Since the GGI was developed on the Affymetrix microarray platform, we used this dataset to compute the corresponding risk group predictions. Note that, in order to lead to equivalent proportions of risk group predictions ($\approx$ 30% of the patients in the low-risk group), we identified a cutoff for the GGI in the VDX dataset.

5.2.2.3 Results

First, we observed from the large scale comparative study of risk prediction models that only the GGI signature outperformed the single proliferation gene AURKA in at least two independent datasets (Tables 3 and 5 in [Haibe-Kains et al., 2008c]). More complex risk prediction models failed to do so, suggesting that the increase in variance due to their complexity is not sufficiently counterbalanced by the decrease in bias to yield good performance in microarray data analysis.

Second, we computed the risk group predictions for the GGI, GENE70 and GENE76 signatures in the TBG dataset. We observed high concordance between risk group predictions computed by the three signatures (Figure 5.3). Note that, while GENE70 and GGI yielded

very similar classifications, GENE76 appeared to deviate more, with 20% of the patients classified differently by this signature.

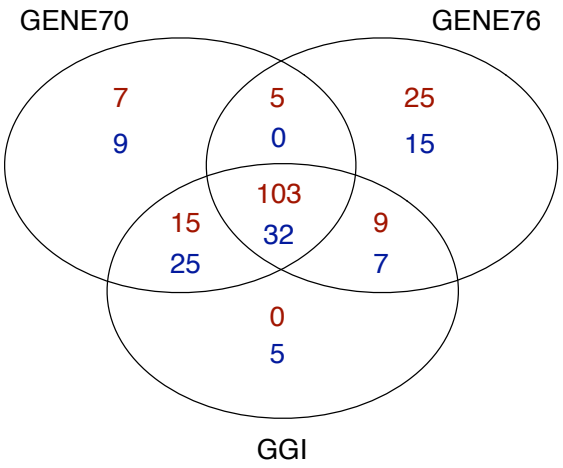


Figure 5.3: Concordance in risk group predictions between GENE70, GENE76 and GGI in the TBG dataset. Red numbers are for the high-risk patients and blue numbers are for the low-risk patients.

In order to assess whether these classification discrepancies have an impact on the prognostic ability of the signatures, we estimated the survival curves of the concordant and discordant cases (Figure 5.4) and we performed a statistical performance comparison using the concordance index and the hazard ratio (Figure 5.5).

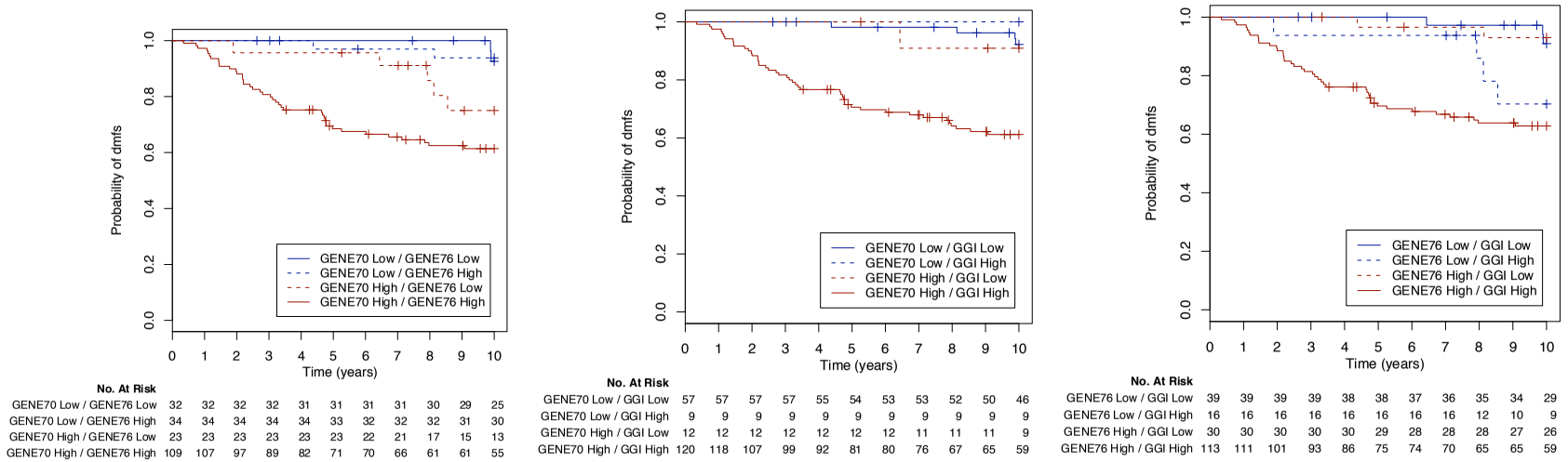
Again we observed in Figure 5.4 (b) that the survival curves of the GENE70 and GGI risk groups are similar, the discordant cases revealing good survival globally. In contrast, the low-risk group of GENE76, including patients classified as high-risk by GENE70 and GGI (Figures 5.4 (a) and (c) for the comparison with GENE70 and GGI respectively), have a good prognosis until six years, but their survival probability drops dramatically afterwards.

We assessed and compared quantitatively the performance of these gene signatures, as well as a prognostic clinical model Adjuvant! Online (AOL), through the novel statistical framework described in Section 4.4. Figures 5.5 (a) and (b) show the forestplot of the concordance indices and the hazard ratio, respectively. Although all the concordance indices of the signatures are highly significant, GENE70 and GGI displayed a higher concordance index than GENE76 (0.90 compared to 0.80; Figure 5.5 (a)). However, this difference was not statistically significant. In contrast, the clinical risk calculated using AOL displayed a lower concordance index (0.69) when compared to either one generated by the gene signatures. GENE70 and GGI yielded a significantly better concordance index than AOL, unlike GENE76. We observed similar results for the hazard ratio in Figure 5.5 (b).

5.2.2.4 Findings

The comparative study of the GGI led to three main findings:

- Complex risk prediction models for breast cancer prognostication do not outperform the simplest risk prediction models in independent datasets. Indeed, in our large scale



	3 yrs	5 yrs	10 yrs
GENE70 Low/GENE76 Low	1.00	1.00	0.93
GENE70 Low/GENE76 High	1.00	0.97	0.94
GENE70 High/GENE76 Low	0.96	0.96	0.75
GENE70 High/GENE76 High	0.80	0.68	0.61

(a)

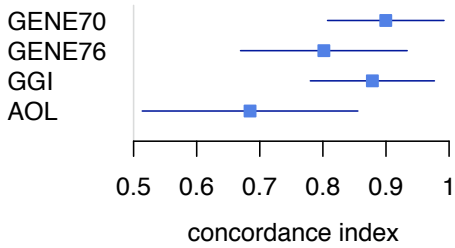
	3 yrs	5 yrs	10 yrs
GENE70 Low/GGI Low	1.00	0.98	0.92
GENE70 Low/GGI High	1.00	1.00	1.00
GENE70 High/GGI Low	1.00	1.00	0.91
GENE70 High/GGI High	0.81	0.70	0.61

(b)

	3 yrs	5 yrs	10 yrs
GENE76 Low/GGI Low	1.00	1.00	0.91
GENE76 Low/GGI High	0.94	0.94	0.70
GENE76 High/GGI Low	1.00	0.97	0.93
GENE76 High/GGI High	0.81	0.69	0.63

(c)

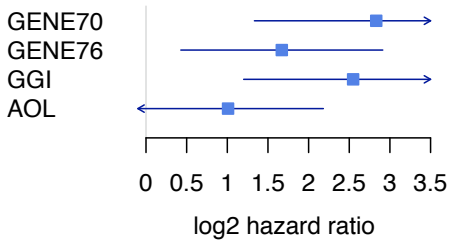
Figure 5.4: Survival curves for: (a) GENE70 vs GENE76; (b) GENE70 vs GGI, and (c) GENE76 vs GGI. The tables report the probability of survival for each stratum at three, five and ten years.



	Test difference
GENE70 vs GENE76	0.15
GENE70 vs GGI	0.53
GENE76 vs GGI	0.22

	Test superiority
GENE70 vs AOL	0.007
GENE76 vs AOL	0.13
GGI vs AOL	0.015

(a)



	Test difference
GENE70 vs GENE76	0.11
GENE70 vs GGI	0.42
GENE76 vs GGI	0.19

	Test superiority
GENE70 vs AOL	0.015
GENE76 vs AOL	0.21
GGI vs AOL	0.026

(b)

Figure 5.5: Statistical performance comparison between GENE70, GENE76, GGI and AOL. Forestplot of the performance of the risk group predictions and tables of p-values form the statistical comparison to test the difference between the performance of the signatures (two-sided test) and to test the superiority of the signatures over AOL (one-sided test): (a) concordance index and (b) $\log_2$ hazard ratio.

comparative study of AURKA, GGI and simple to complex risk prediction models, we observed that complex methods performed very well in the training set. However, their performances in the independent datasets were poorer, and they failed to outperform consistently the simplest model, i.e. AURKA. These results highlight the fact that the loss of interpretability deriving from the use of overly complex methods in survival analysis of breast cancer microarray data might be not sufficiently counterbalanced by an improvement in the quality of prediction.

- Simple, yet robust, quantification of proliferation at the molecular level yields good performance for breast cancer prognostication. Interestingly, AURKA, the simplest model, defining the risk score as the expression of a single proliferation gene, performed well in all the prognostic tasks. The GGI was the only model that outperformed AURKA in at least two independent datasets, whatever the performance criterion for risk score and risk group predictions. As the GGI is a signed average of expressions of proliferation related genes, these results highlight that simple, yet robust, risk prediction models yield similar or even better performance than complex ones. Moreover, these results highlight the importance of proliferation measured by gene expression profiling in breast cancer prognostication, and confirm the results of [Sotiriou et al., 2006].

- The GGI is competitive with state-of-the-art prognostic gene signatures that include many genes related to other biological processes than cell proliferation. Indeed, we compared the prognostic ability of the GGI and two state-of-the-art gene signatures, namely GENE70 and GENE76, as well as the clinical model Adjuvant! Online (AOL). Interestingly, we observed that the GGI is competitive with the more complex GENE70 and GENE76 signatures and is significantly better than AOL, the prognostic clinical model. Since the GGI is driven by proliferation-related genes, these results suggest that proliferation might be the driving force of the GENE70 and GENE76 signatures. We will consider this question in greater detail in Section 5.4.1.

5.2.3 Tamoxifen Resistance Signature (TAMR13)

5.2.3.1 Motivations

We have seen in Section 3.3.1 that the prediction outcome of patients treated by tamoxifen, a widely used anti-estrogen therapy, is the subject of intense research [Ma et al., 2004; Paik et al., 2004]. However, there are only few biomarkers routinely used that can predict response to commonly prescribed therapies. The presence of estrogen receptors (ER) is the best indicator of response to anti-estrogen agents such as tamoxifen. However, 30-40% of women with ER-positive breast cancer will develop distant metastases and die despite tamoxifen treatment. The underlying biological mechanisms of resistance to tamoxifen are incompletely understood.

We aimed to use microarray technology to build a predictive model for the resistance to tamoxifen therapy. The purpose of the model is ultimately to facilitate our understanding of the biological underpinnings of resistance mechanisms

The experimental findings of this study as presented below were published in [Loi et al., 2008], and the methodology was the subject of a book chapter [Haibe-Kains et al., 2008a]. Note that the present work complies with the *research reproducibility* guidelines proposed

in [Gentleman, 2005] regarding the availability of the code and the reproducibility of results and figures[1].

5.2.3.2 Methods

In order to extract a gene signature predictive of tamoxifen resistance, we used the methodology presented in Section 4.1.

1. Feature transformation: We used an independent dataset of primary breast tumors from untreated patients (UNT dataset, see Table 5.1) to apply the procedure described in Algorithm 2 with the hyperparameters $h = 0.5$ and $s = 5$. From this clustering model, we applied the procedure described in Algorithm 3 to compute the features, called *pclust*, on the dataset of early breast tumors from tamoxifen treated patients (TAM dataset, see Table 5.1). Note that the gene expressions of tumors from the untreated and tamoxifen treated patients are comparable, i.e. early breast tumors before any treatment.

2. Feature selection: The features are ranked with respect to their prognostic value. The scoring function in the feature ranking procedure described in Algorithm 4 is defined as the significance of the hazard ratio (Section 3.4.1). The number of genes in the signature is selected by maximizing the stability through the use of the *Stab* criterion introduced in Section 4.1.2.1.

3. Model building: Once the signature was identified, we built a predictive model as described in Section 4.1.3.1, except that the coefficients β in Equation (4.5) were estimated through the univariate Cox models fitted during the feature ranking step.

5.2.3.3 Results

Gene signature As a training set, we used 255 tamoxifen treated patients from the TAM dataset (see table 5.1). After the feature transformation, 110 features (pclust) remained to perform the feature ranking. A signature of 13 pclusts was assessed to be highly stable (Figure 5.6) and hence chosen to build the predictive model, denoted by TAMR13 hereafter. Figure 5.7 shows the frequency of selection of each pclust in the sampling process.

All of the 13 clusters incorporated in the final model are the most frequently selected during the training phase. The list of genes included in each of the 13 clusters includes 181 unique genes and is referred to as the TAMR13 signature (Appendix C.2).

The biological functions of each of the 13 clusters were analyzed by using Ingenuity Pathway Analysis (IPA; [Ingenuity Systems][2]). Table 4 in the paper lists the high level functions and associated canonical pathways, with statistically significant enrichment for each cluster. There are several gene clusters related to cell cycle function, supporting the fact that the GGI was shown to be also associated with tamoxifen resistance [Loi et al., 2007].

[1] Raw gene expression and clinical data are publicly available in the GEO public database [Barrett et al., 2005] and the Sweave version of the book chapter including the standalone R code [R Development Core Team, 2007] is available at `http://www.ulb.ac.be/di/map/bhaibeka/cichapter/`.

[2] IPA is an all-in-one software application that enables researchers to model, analyze, and understand the complex biological and chemical systems at the core of life science research. This tool is widely used to interpret gene signatures.

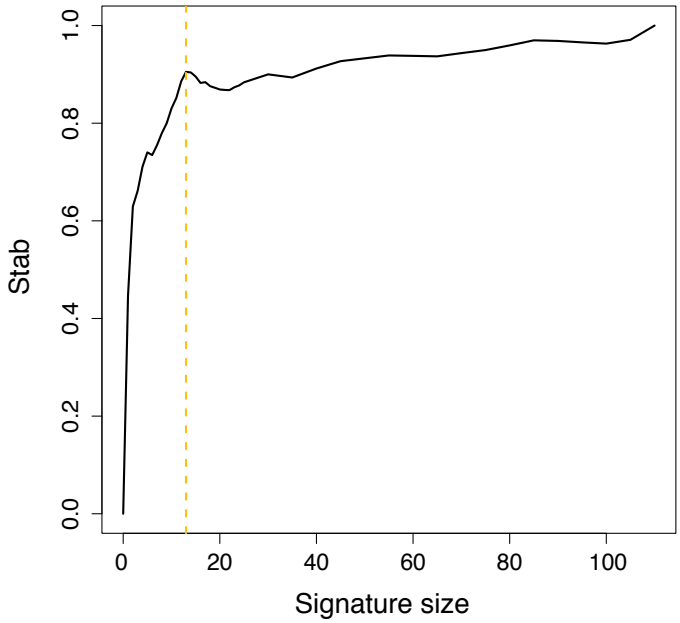


Figure 5.6: Stability of the signature with respect to the size. The vertical orange dashed line represents the size selected as a good trade-off between stability and signature size. Note that the stability converges to 1 with increasing size since the *Stab* criterion was used instead of the $Stab_{adj}$ criterion in [Loi et al., 2008].

Pclust 110 contains genes that have previously been associated with chemotaxis and invasion of breast cancer cell lines (SLIT2, RECK) [Liu et al., 2003; Prasad et al., 2004], as well as genes related to the extracellular matrix (ECM2, COL4A1). Less well characterized is the role of lipid metabolism (pclust 79) and immunological aspects in the differential response to tamoxifen (pclust 784 and pclust 375), though TNF alpha and TGF beta have previously been implicated in breast cancer development and progression [Turner et al., 2004]. A functional analysis of pclust 375 suggests that these genes (TGFBR4, PTGER4, C3, GNG2) are mainly involved in cellular inflammatory response and could be particularly important in determining the host's response to tamoxifen. The presence of gene clusters in TAMR13 that allude to other biological pathways apart from cell cycle function may facilitate our further understanding of the upstream mechanisms behind tamoxifen resistance.

Performance assessment We assessed the performance of our model by estimating the hazard ratio on several independent datasets, using DMFS as survival endpoint. The main independent dataset was a set of 77 patients from the TAM dataset (same consecutive series of patients as in the training set). This dataset is referred to as GUYT2. The survival curves of the low and high-risk groups of patients are presented in Figure 5.8. The two survival curves are significantly different (logrank test p-value < 0.03) and the hazard ratio of the risk group predictions is large ($HR = 4.02$, 95%CI $[1.13, 14.27]$).

We further validated our model on two additional independent datasets of tamoxifen treated patients, specifically MGH (see Table 5.1) and a private dataset from [Reid et al., 2005]. We computed the hazard ratio in each dataset separately and combined the estima-

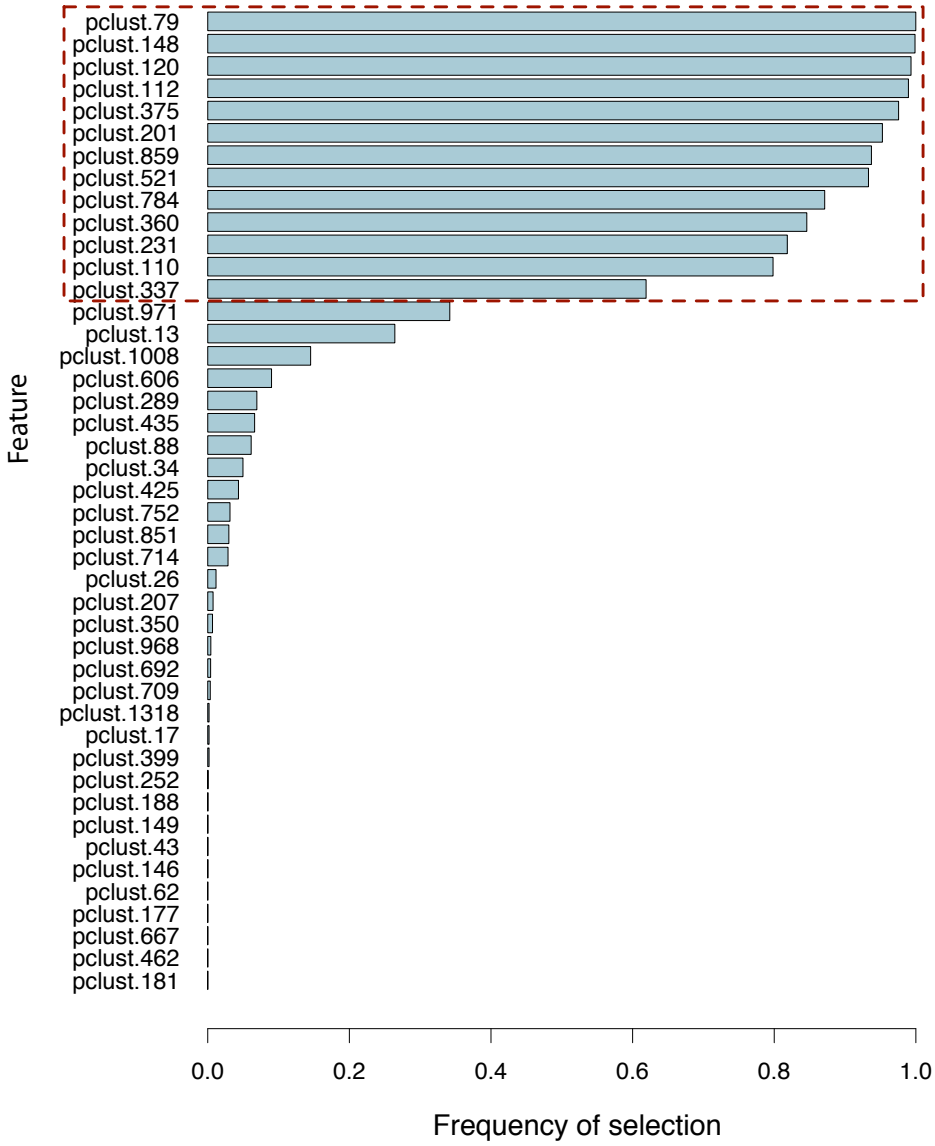


Figure 5.7: Frequency of selection of the most frequently selected features during the signature stability assessment. The red box delimits the set of the 13 most frequently selected pclusts.

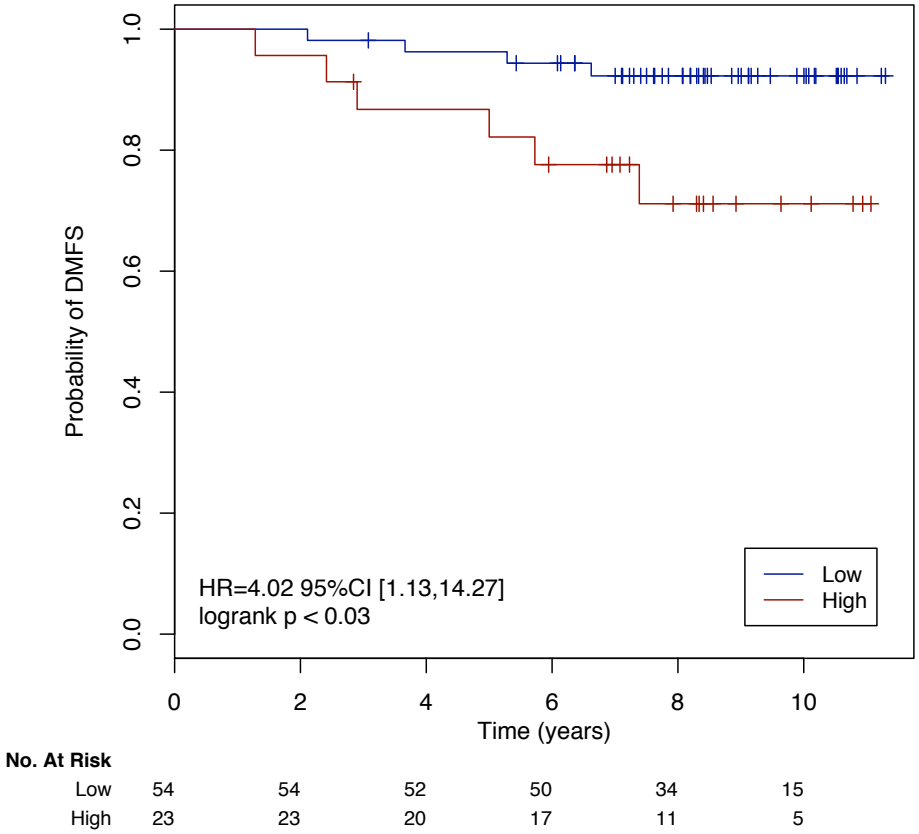
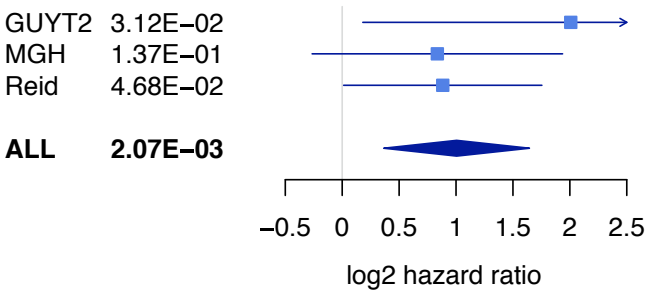


Figure 5.8: Survival curves of the risk group predictions in the independent dataset GUYT2.



Data set	Hazard Ratio	95%CI	P value
GUYT2	4.02	(1.13-14.27)	0.03
MGH	1.78	(0.83-3.83)	0.1
Reid	1.84	(1.01-3.37)	0.05
ALL	2.01	(1.29-3.13)	0.002
Test for heterogeneity p=0.5			

Figure 5.9: Performance assessment of the risk group predictions in the three independent datasets GUYT2, MGH and Reid. The hazard ratio estimates are combined to get an overall estimate of the performance of our model (triangle).

tions using the inverse variance-weighted method with fixed effect model[3] [Cochrane, 1954]. The overall hazard ratio is equal to 2.01 and is significant (Wald test p-value < 0.002). The survival curves of the risk group predictions in the MGH and Reid datasets have been published in the additional file 6 of [Loi et al., 2008].

5.2.3.4 Findings

In this study, we have developed a gene signature (TAMR13) predictive of the resistance to tamoxifen for ER-positve breast cancer patients. The approach we used to identify the gene signature facilitates both signature stability and biological interpretation. These are critical issues in the challenging task of building prognostic gene signatures for breast cancer patients as we endeavor to derive biologically meaningful and clinically useful information from microarray data.

The main findings of this study are twofold:

- From a biological point of view, while our study emphasized the important role of proliferation genes in prognosis, we showed that our signature includes other genes and pathways that may elucidate further mechanisms that influence clinical outcome and prediction of resistance to tamoxifen.
- From a prognostic point of view, we showed in several independent datasets that the gene signature was able to distinguish patients at high risk of distant metastasis despite adjuvant tamoxifen monotherapy. These poor prognosis patients could be selected for prescription of other treatment modalities, such as chemotherapy and/or biological agents.

While in the future we may have a microarray-based diagnostic test incorporating all 181 genes in the 13 clusters, at present the routine use of this technology is not logistically feasible. However, the advantage of our approach is that, because each cluster consists of a group of highly correlated genes, the clusters can effectively act as one covariate. Thus, a diagnostic test of just 13 genes (one per cluster) could be developed for clinical use if desired, even though for biological research one would be more interested in all the genes per cluster.

5.3 Breast Cancer Molecular Subtypes

5.3.1 Motivations

In the previous section we showed that we were able to identify prognostic gene signatures for the global population of breast cancer patients, and that these signatures yielded good performance. We also mentioned in Section 3.2 that early microarray studies highlighted the existence of different breast cancer molecular subtypes. The next natural step was then to study the relationship between the prognostic gene signatures and the breast cancer molecular subtypes, the hope being to combine them in order to improve the current prognostic

[3]The choice of the fixed effect model instead of the random effect model is driven by the test of heterogeneity between the hazard ratios estimated in each dataset separately. Since the estimates were not significantly heterogeneous, we used the fixed effect model [Cochrane, 1954].

models. Therefore, we introduced a novel subtype clustering model to accurately classify tumors according to their subtypes (Section 4.2), and we used this model to identify local prognostic gene signatures (Section 4.3).

In this section, we will present the experimental findings related to identifying breast cancer molecular subtypes. Specifically, we will demonstrate the robustness of our novel subtype clustering model through its validation in numerous independent datasets and its comparison with the state-of-the-art techniques. Some of these results have been published in [Desmedt et al., 2008; Haibe-Kains et al., 2009].

5.3.2 Methods

In order to build our subtype clustering model, we followed the flowchart depicted in Figure 4.5:

1. Gene clustering: The ESR1 and ERBB2 module scores were computed for each tumor using our prototype-based feature transformation method (Section 4.2.1).
2. Patient clustering: The subtype clustering model was fitted using these two gene module scores (Section 4.2.2).

These two steps are described in the following paragraphs.

Prototype-based feature transformation At this step, we aimed to identify sets of genes, called gene modules, to quantify the activity of the key biological processes in breast cancer described in Section 3.1. Such biological processes include the ER and the HER2 signaling pathways, relevant for the identification of breast cancer molecular subtypes as shown in [Perou et al., 2000; Sorlie et al., 2001, 2003; Sotiriou et al., 2003; Hu et al., 2006; Kapp et al., 2006]. Since we will present the experimental findings from a meta-analysis of the prognostic value of the key biological processes according to the breast cancer molecular subtypes in Section 5.4.1, we describe hereafter the identification of a gene module for each biological process described in Section 3.1.

We selected a prototype gene for each biological process:

- The prototype gene ESR1 represents the ER signaling pathway.
- The prototype gene ERBB2 represents the HER2 signaling pathway.
- The prototype gene AURKA represents the proliferation.
- The prototype gene PLAU represents the tumor invasion.
- The prototype gene VEGF represents the angiogenesis.
- The prototype gene STAT1 represents the immune response.
- The prototype gene CASP3 represents the apoptosis.

These 7 prototypes, represented in Figure 5.10, were used to populate the corresponding gene modules with genes that are specifically co-expressed with them (Section 4.2.1). Each gene module was then summarized by computing the weighted average of the expression of

the specific genes (Algorithm 3). As a feature transformation method, it reduced the original matrix of expressions to a matrix of 7 dimensions defined by the gene module scores, including the ESR1 and ERBB2 module scores used in the subtype clustering model presented below.

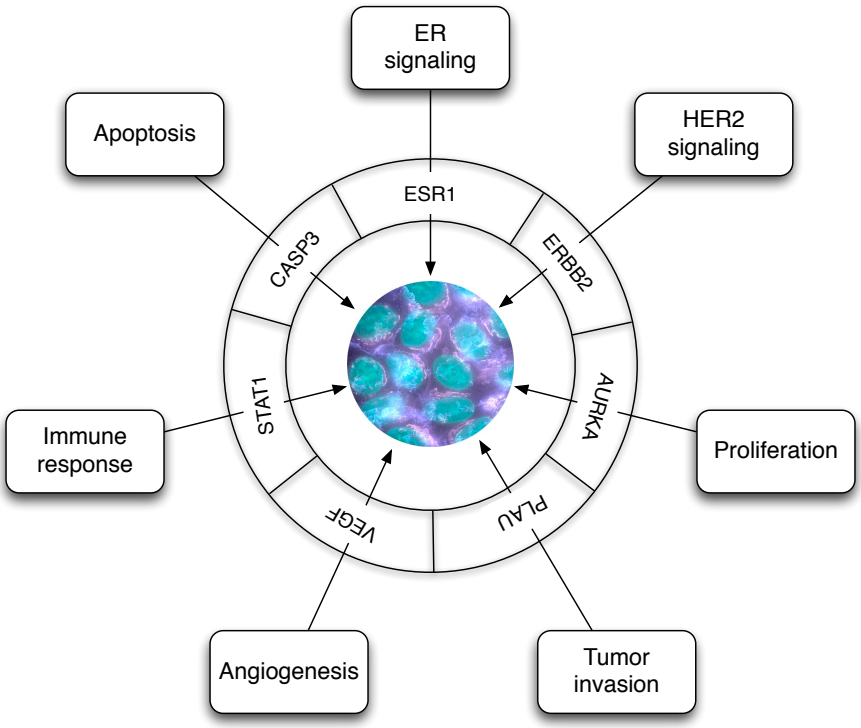


Figure 5.10: Illustration of the key biological processes involved in breast cancer (boxes) with their corresponding prototype genes (gene names in the ring).

Each gene module score was rescaled to get comparable scales between datasets. Indeed, the use of different microarray platforms and different normalization methods may lead to different scales for gene expressions and, consequently, for gene module scores. The gene module scores were therefore scaled such that quantiles 2.5% and 97.5% are equaled to -1 and $+1$ respectively. This scaling is robust to outliers and ensured that the scores lay approximately in $[-1, +1]$.

Subtype clustering We used the ESR1 and ERBB2 module scores to fit the subtype clustering model described in Section 4.2.2 to a training dataset. We then assessed the performance of the model in the independent datasets. To do so, we used the prediction strength described in Section 2.2.4. The idea was to view the clustering analysis (unsupervised learning, see Section 2.1.3.3) as a supervised classification problem (supervised learning, see Section 2.1.3.3) in which the "true" class labels have to be estimated. To assess the performance of a clustering model in an independent dataset, we first fitted a new clustering model to this dataset, using the same method used for the original clustering model to estimate the "true" class labels. Second, we compared the "true" class labels with the class labels returned by the original clustering model, through prediction strength [Tibshirani and

Walther, 2005]. If prediction strength was close to 1, it meant that our clustering model fitted the independent data well, since the two clustering models returned the same class labels; otherwise, prediction strength was close to 0.

5.3.3 Results

Gene modules We used the two largest datasets of global populations of breast cancer patients, i.e. VDX and NKI, to perform the prototype-based feature transformation. VDX and NKI were generated from two different microarray technologies (Table 5.1), Affymetrix and Agilent respectively, with $\approx 10,000$ genes in common. We combined these two datasets through a meta-analytical framework described in the Supplementary Information section of the paper [Desmedt et al., 2008]. We did not attempt to optimize the hyperparameters of the method and used the values $c = 0.05$ and $e = 0.95$ to identify modules of a size larger than five genes each. The resulting number of specific genes in each module is reported in Table 5.2.

Gene module	Size
ESR1	469
ERBB2	28
AURKA	229
PLAU	68
VEGF	14
STAT1	95
CASP3	9

Table 5.2: Number of genes specifically co-expressed with the prototype in the corresponding module. The size includes the prototype itself.

The large size of the ESR1 and AURKA modules highlights their broad effect on the gene expression profile of the breast tumors. Interestingly, the ERBB2 module contained only a reasonable number of specific genes, despite having been shown in previous publications that these genes are of the highest importance for breast cancer subtyping [Perou et al., 2000; Sorlie et al., 2001, 2003; Sotiriou et al., 2003; Hu et al., 2006; Kapp et al., 2006]. This suggests that the expressions of the genes in this module have a strong impact on clustering analysis used to identify the breast cancer molecular subtypes. The list of genes for each module is given in Appendix C.3.

The 7 gene modules were analyzed by using ingenuity pathway analysis [Ingenuity Systems]. The ESR1 module was composed of 469 genes and, as expected, was characterized by the co-expression of numerous luminal and basal genes already reported in previous microarray studies such as XBP1, TFF1, TFF3, MYB, GATA3, PGR and several keratins. The ERBB2 module included 28 genes, with nearly half co-located on the 17q11-22 amplicon, such as THRA, ITGA3 and PNMT. The proliferation module (AURKA) comprised 229 genes, with 34 of them represented in the previously reported GGI [Sotiriou et al., 2006]. The majority of these genes, such as CCNB1, CCNB2, BIRC5, were involved in cellular growth and proliferation, and cancer and cell cycle related functions. The tumor invasion/metastasis module (PLAU) included 68 genes with several metalloproteinases among them. These

genes were significantly associated with functions such as cellular movement, tissue development, cellular development and cancer related functions. The immune response module (STAT1) was made up of 95 genes, the majority being associated with immune response, followed by cellular growth and proliferation, cell-signaling and cell death. The angiogenesis module (VEGF) consisted of 10 genes related to cancer, gene expression, lipid metabolism and small molecule biochemistry. Finally, the apoptosis module (CASP3) only included 9 genes mainly associated with protein synthesis and degradation, as well as cellular assembly and movement.

Subtype clustering model We chose VDX as a training set since this dataset includes a large number of experiments on the Affymetrix platform. The use of NKI as a training dataset instead of VDX led to a virtually identical model (data not shown), but had the drawback of using Agilent microarray technology, which is less widely used (Table 5.1).

In order to identify the most likely number of clusters (subtypes) present in the data, we fitted the subtype clustering model with increasing number of clusters (one to ten) and then computed the corresponding BIC values (Section 2.2.2.1). Since the BIC values are dependent on sample size and data distribution, we scaled the BIC values such that the value for a single cluster was equal to zero, and the dispersion between the maximum and the minimum values was equal to unity. This scaling procedure made it possible to compare BIC values between datasets.

It is apparent in Figure 5.11 that the BIC estimates increased dramatically until three clusters and reached a plateau afterwards. Therefore, we considered a mixture of three Gaussians, since this number of clusters is likely given the data. We referred to these clusters as the ER-/HER2-, HER2+, and ER+/HER2- subtypes with respect to their typical values for the ESR1 and ERBB2 module scores.

The parameters of the subtype clustering model using three Gaussians are given in Table 5.3. As mentioned in Section 4.2.2, the covariance structure is constrained such that the covariance matrices of the Gaussians are diagonal and equal, as in Equation (4.11).

	ER-/HER2-	HER2+	ER+/HER2-
μ			
ESR1	-0.77	0.09	0.59
ERBB2	-0.71	0.68	-0.26
Σ			
ESR1		0.062	
ERBB2		0.063	
π	0.29	0.16	0.55

Table 5.3: Subtype clustering model: parameters of the mixture of three Gaussians fitted on the training dataset (VDX).

Figure 5.12 sketches the density distribution of such a mixture of three Gaussians.

The scatterplot in Figure 5.13 illustrates the classification of the tumors by the subtype clustering model on the training dataset (VDX), each subtype being represented by a different color and symbol. The tumors are classified with respect to their maximum posterior probability computed by the subtype clustering model as in Equation (2.7).

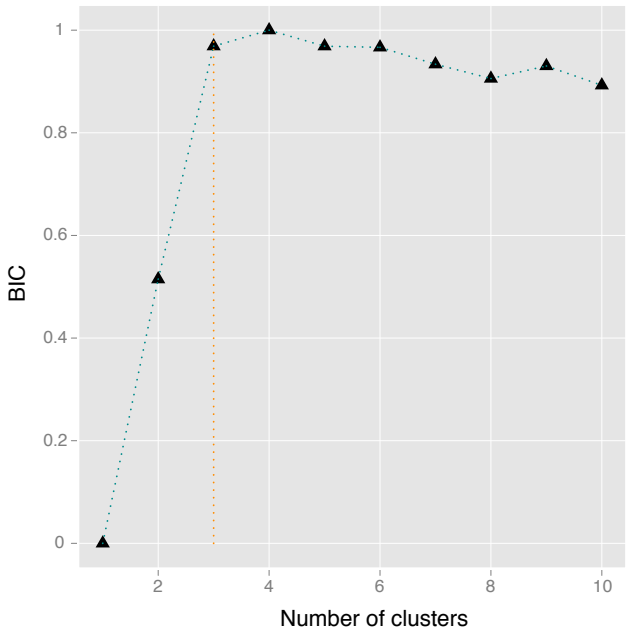


Figure 5.11: Evolution of the scaled BIC estimates of the subtype clustering with respect to the number of clusters in the training dataset (VDX). The vertical orange dashed line represents the number of Gaussians selected for the subtype clustering model.

One of the advantages of model-based clustering is that it can easily predict the class of new data (Section 4.2.2). This allows for performance assessment of the model in independent datasets.

Performance assessment We used the subtype clustering model in all the datasets described in Table 5.1, which represent a global population of breast cancer patients, excluding MGH, TAM and TAM2. We first identified the breast tumor subtypes and then computed the prediction strength of the model in each independent dataset.

One can see in Figures B.1-B.5 (Appendix B.1) that the pattern of the three subtypes observed in the training dataset (VDX, see Figure 5.13) was well preserved over the independent datasets, except for MUG, in which the subtypes were not easily discriminated.

The prediction strength of the subtype clustering model applied to each independent dataset is reported in Table 5.4. The prediction strengths for the ER-/HER2-, HER2+, and ER+/HER2- subtypes separately are shown in the first columns, and the overall prediction strengths, denoted by *ps*, are given in the last column. The last two rows report the mean and the standard deviation of the prediction strengths. Note that, in [Tibshirani and Walther, 2005], the authors stated that a prediction strength is considered to be "good" if $ps \geq 0.8$.

We observed good overall prediction strengths for most of the datasets, except for STNO2, DUKE2 and MUG. However the mean *ps* is equal to 0.83 ± 0.12, highlighting the good performance of the subtype clustering model applied to new data. Looking at the prediction strengths for each subtype separately, we observed that the ER-/HER2- was particularly

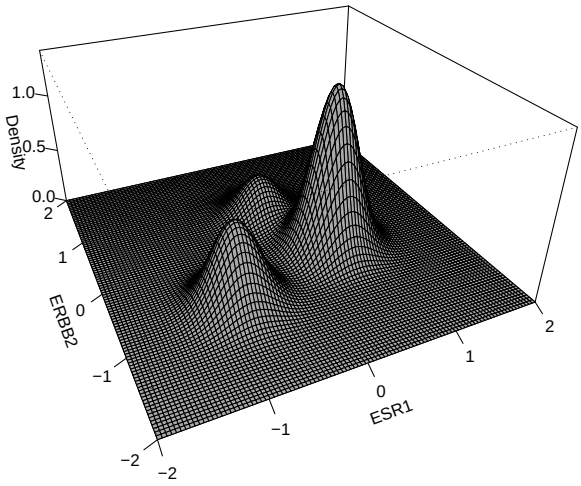


Figure 5.12: Density distribution of the mixture of three Gaussians fitted for the subtype clustering model.

Dataset	ER-/HER2-	HER2+	ER+/HER2-	*ps*
NKI	1.00	1.00	1.00	1.00
TBG	1.00	1.00	0.83	0.83
UPP	1.00	0.93	0.87	0.87
UNT	1.00	0.89	0.92	0.89
STNO2	1.00	0.69	0.97	0.69
NCI	0.85	0.83	0.93	0.83
STK	1.00	0.91	0.87	0.87
MSK	1.00	1.00	0.96	0.96
UNC2	1.00	0.87	0.96	0.87
NCH	1.00	0.82	0.98	0.82
DUKE	1.00	0.82	0.92	0.82
DUKE2	1.00	0.64	0.95	0.64
MAINZ	1.00	1.00	0.90	0.90
CAL	1.00	1.00	0.95	0.95
LUND2	1.00	0.89	0.87	0.87
LUND	1.00	1.00	0.81	0.81
MUG	0.66	0.61	0.49	0.49
mean	0.97	0.88	0.89	0.83
sd	0.09	0.13	0.12	0.12

Table 5.4: Prediction strength of the subtype clustering model in each independent dataset. The mean and the standard deviation of the prediction strengths are reported in the last two rows.

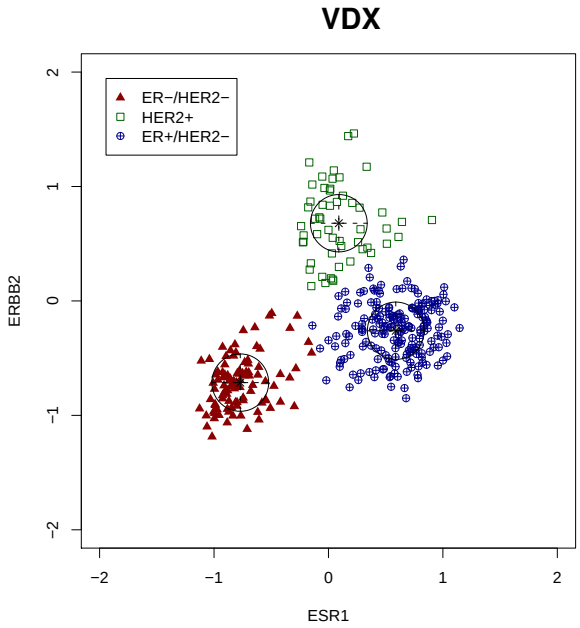


Figure 5.13: Tumors in the training dataset (VDX) colored by their subtype as defined by their maximum posterior probability computed by the subtype clustering model. Each subtype is represented by a different color and symbol. The superimposed ellipses correspond to the covariance of the components.

well identified (mean *ps* equal to 0.97 ± 0.09), while the HER2+ and ER+/HER2- yielded similar prediction strengths (mean *ps* equal to ≈ 0.88).

In order to assess whether the number of clusters selected in the training dataset (VDX) was likely given the independent datasets, we computed the BIC values for the clustering model using one to ten clusters in each independent dataset. The evolution of the mean BIC estimates with respect to the number of clusters is given in Figure 5.14. We observed that three is the most likely number of clusters, supporting our choice made for the training dataset.

Subtypes and clinical outcome Initial studies [Perou et al., 2000; Sorlie et al., 2001, 2003; Sotiriou et al., 2003] showed that patients exhibit different clinical outcomes depending on the molecular subtypes of their breast tumors (Section 3.2). However, due to the small sample sizes of these studies, the authors were only able to study heterogeneous populations of patients (e.g. different treatments or different stages of the disease). Therefore, the conclusions about clinical outcome might be misleading due these potentially confounding factors.

For this thesis, we collected numerous breast cancer microarray datasets (Table 5.1). This allowed us to study a homogenous population of interest, i.e. untreated patients with early (node-negative) primary breast tumors. In the independent datasets, we retrieved these patients from the NKI, TBG, UPP, UNT and MAINZ datasets (745 patients). Their survival curves using DMFS as survival endpoint (except for UPP for which only RFS is

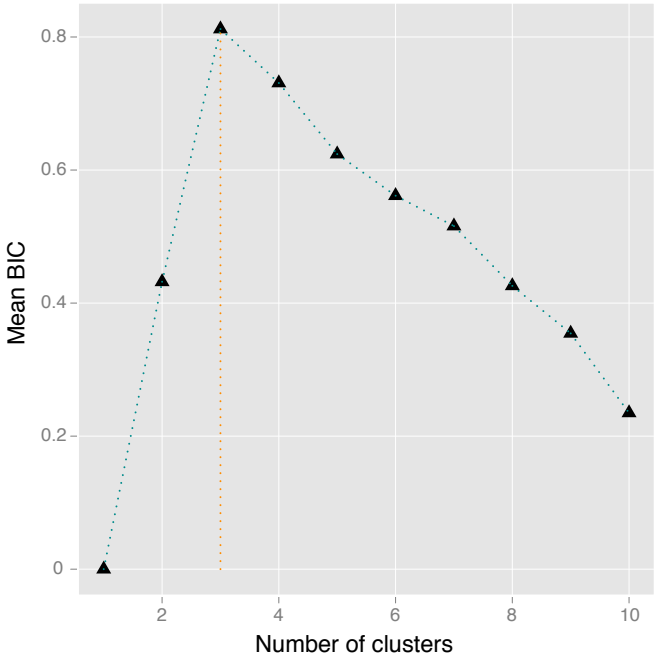


Figure 5.14: Evolution of the mean BIC values estimated from each independent dataset, with respect to the number of clusters.

available) and stratified by molecular subtypes, are shown in Figure 5.15.

We observed a significant difference between the survival curves (logrank test p-value of 4E-6), the patients who have a ER+/HER2- tumor exhibiting a better survival than the patients who have a ER-/HER2- or HER2+ tumor. The ER-/HER2- and HER2+ subtypes exhibit similar survival. The ER+/HER2- subtype, including 69% of the tumors, exhibits better survival than the ER-/HER2- and HER2+ subtypes, including 16% and 15% of the tumors respectively. These results reinforce the observations made in the initial publications.

Performance comparison: gene modules vs prototypes We assessed here whether the subtype clustering model using the ESR1 and ERBB2 module scores yields better performance than a clustering model using only their prototypes (the expressions of the single genes ESR1 and ERBB2). To do so, we fitted the same clustering model (mixture of three Gaussians) onto the training dataset (VDX), but used the gene expressions of the ESR1 and ERBB2 prototypes instead of their gene module scores. We computed the prediction strengths just as we did previously (Table 5.7). We observed globally lower prediction strengths and larger deviations for each subtype separately and for the overall *ps*. In particular, *ps* is equal to 0.75 ± 0.25, which does not fulfill the condition for "good" prediction strength as defined in [Tibshirani and Walther, 2005]. Note the presence of extremely low prediction strengths for some datasets: *ps* = 0 for NCH and *ps* = 0.40 for MAINZ. In contrast, we obtained *ps* = 0.82 for NCH and *ps* = 0.90 for MAINZ using the subtype clustering model with gene module scores.

Although using prototypes instead of the gene module scores globally decreased the

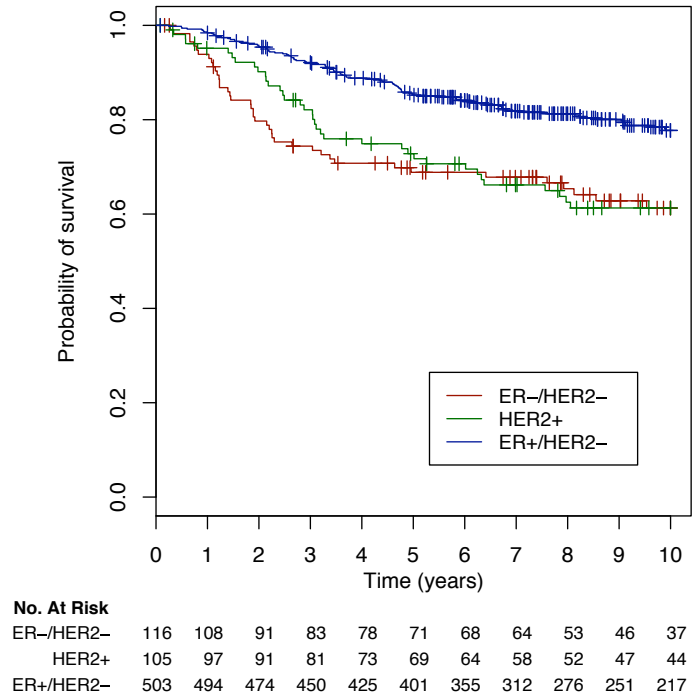


Figure 5.15: Survival of untreated node-negative breast cancer patients with respect to their tumor subtypes. The patients come from the NKI, TBG, UPP, UNT and MAINZ datasets.

Dataset	ER-/HER2-	HER2+	ER+/HER2-	*ps*
NKI	1.00	0.96	0.95	0.95
TBG	1.00	1.00	0.95	0.95
UPP	1.00	0.94	0.99	0.94
UNT	1.00	1.00	0.98	0.98
STNO2	0.83	1.00	0.82	0.82
NCI	0.66	1.00	1.00	0.66
STK	1.00	0.81	0.95	0.81
MSK	1.00	1.00	0.96	0.96
UNC2	1.00	1.00	0.86	0.86
NCH	0.49	0.92	0.00	0.00
DUKE	0.70	0.76	1.00	0.70
DUKE2	0.86	0.81	1.00	0.81
MAINZ	0.60	0.40	0.93	0.40
CAL	1.00	1.00	1.00	1.00
LUND2	0.74	1.00	0.92	0.74
LUND	0.75	1.00	0.93	0.75
MUG	0.49	0.67	0.77	0.49
mean	0.83	0.90	0.88	0.75
sd	0.19	0.16	0.24	0.26

Table 5.5: Prediction strength of the subtype clustering model using the prototypes in each independent dataset. The mean and the standard deviation of the prediction strengths are reported in the last two rows.

prediction strength of the subtype clustering model, the superiority of the model using the gene module scores was not significant (Wilcoxon Rank Sum test p-value of 0.26).

Performance comparison with Perou's method We also compared the performance of our subtype clustering model with the method introduced in [Perou et al., 2000] and further used in [Sorlie et al., 2001, 2003; Hu et al., 2006]. To do so, we used the procedure described in Section 3.2.

Perou's clustering model We applied Perou's method to identify three subtypes in the training dataset (VDX). This method revealed similar subtypes to those identified by our novel subtype clustering model (Table 5.6). The association between the two classifications is strong (Cramer's *V* statistic of 0.84, [Cramer, 1999]) and significant (χ^2 test p-value $<$ 1E-16, [Plackett, 1983]).

Novel method	Perou's method		
	cluster 1	cluster 2	cluster 3
ER-/HER2-	97	2	0
HER2+	0	46	8
ER+/HER2-	5	16	170

Table 5.6: Contingency table to assess the concordance between the subtype identification of our novel method and Perou's method.

Performance assessment We appled this model to the same independent datasets as before in order to compare its performance with our model. To do so, we computed the prediction strengths reported in Table 5.7. We observed globally lower prediction strengths for each subtype separately and for the overall *ps*. Specifically, *ps* was equal to 0.47 ± 0.10 and therefore does not fulfill the condition for a "good" prediction strength as defined in [Tibshirani and Walther, 2005]. Note the presence of low prediction strengths for some datasets, e.g. *ps* = 0.33 for MUG. The superiority of our subtype clustering model is highly significant (Wilcoxon Rank Sum test p-value of 8E-6).

When we applied Perou's method to identify four to five clusters, as was reported in the original publications, the prediction strengths were even worse (Table B.1.1 in Appendix B.1.1). In contrast, identifying the two main clusters yielded better prediction strengths but, even in this case, the performance was significantly worse than that of our method (Wilcoxon Rank Sum test p-value of 0.044).

5.3.4 Findings

Thanks to the novel methods we developed in this thesis (Section 4.2), we significantly improved the identification of breast cancer molecular subtypes when compared to the state-of-the-art:

- Our subtype clustering model yielded good performance in numerous datasets using different microarray technologies and normalization techniques. We also showed that

	cluster 1	cluster 2	cluster 3	ps
NKI	0.84	0.42	0.60	0.42
TBG	0.92	0.39	0.87	0.39
UPP	0.51	0.61	0.54	0.51
UNT	0.55	0.59	0.55	0.55
STNO2	0.81	0.44	0.78	0.44
NCI	1.00	0.44	0.52	0.44
STK	0.46	0.38	0.50	0.38
MSK	0.89	0.66	0.71	0.66
UNC2	0.93	0.57	0.80	0.57
NCH	0.49	0.57	0.56	0.49
DUKE	0.61	0.42	0.89	0.42
DUKE2	0.97	0.63	1.00	0.63
MAINZ	0.49	0.39	0.60	0.39
CAL	1.00	0.41	0.80	0.41
NCH	0.49	0.57	0.56	0.49
LUND2	0.93	0.51	0.78	0.51
LUND	0.39	0.36	0.49	0.36
MUG	0.42	0.33	0.35	0.33
mean	0.72	0.48	0.67	0.47
sd	0.23	0.11	0.18	0.10

Table 5.7: Prediction strength of the clustering model as fitted by Perou's method in each independent dataset. The mean and the standard deviation of the prediction strengths are reported in the last two rows.

the use of gene module scores instead of single prototype genes yields more stable classifications.

- Our subtype clustering model yielded significantly better performance than Perou's method, the most widely used method in the literature.
- We confirmed in a homogeneous population of untreated node-negative breast cancer patients that the molecular subtypes have different natural histories, ER-/HER2- and HER2+ subtypes having a worse clinical outcome than ER+/HER2- subtype. The accurate classification of breast tumors according to their subtypes is therefore highly important for breast cancer management [Pusztai et al., 2006]. Indeed, targeted therapies are available that are only effective in some subtypes of tumors. For instance, the sensitivity of different molecular subtypes to chemotherapy varies, with ER-/HER2- and HER2+ tumors being more sensitive to chemotherapy [Rouzier et al., 2005] than others.

We will see in Section 5.4.2 how to use the estimation of the posterior probability of belonging to a subtype as computed by our model, in order to identify efficient local prognostic gene signatures.

5.4 Local Prognostic Gene Signatures

In this section, we will present the local prognostic gene signatures we developed and their corresponding experimental findings. These local prognostic genes signatures are closely

related to the breast cancer molecular subtypes (see Section 3.2) and the novel subtype clustering we developed (Sections 4.2.2 and 5.3 for the description of the subtype clustering model and the corresponding experimental findings, respectively).

The experimental results generated from the use of gene modules (Section 5.3) to study the association between known biological processes and breast cancer molecular subtypes from a prognostic point of view will be the subject of Section 5.4.1. The experimental results of the novel prognostic modular model (GENIUS, Section 4.3) will be presented in Section 5.4.2.

5.4.1 Gene Modules and Breast Cancer Molecular Subtypes

5.4.1.1 Motivations

The prognostic relationship between the key biological processes in breast cancer (see [Hanahan and Weinberg, 2000] for a review) and the molecular subtypes was barely known at the time this thesis was begun (Section 3.3.1). Extending the results of the GGI signature, capturing mainly proliferation at the molecular level (Section 5.2.1), we sought to build modules of genes (Section 5.3) in order to quantify the activity of key biological processes in breast cancer and to link them to prognosis with respect to the molecular subtypes.

Beyond the study of the key biological processes, we focussed our research on the performance of the existing prognostic gene signatures with respect to the breast cancer molecular subtypes. Since these gene signatures contain a large number of genes for which the biological function is unknown, it is difficult to discern the driving force behind such signatures from a prognostic point of view.

We collected numerous microarray datasets (Section 5.1) to address these two issues. The experimental results presented below have been published in [Wirapati et al., 2008; Desmedt et al., 2008].

5.4.1.2 Methods

To study the prognostic relationship between the key biological processes in breast cancer and the molecular subtypes, we used the gene modules and the subtype clustering model described in Section 5.3. Accordingly, the subtype clustering model was used for hard partitioning (Section 2.2), the tumors being classified by their maximum posterior probability to belong to a subtype as computed by the model.

In this study, we also analyzed state-of-the-art prognostic gene signatures. Since we were not able to use the original algorithms for all these gene signatures in order to compute them in all the datasets, this being due to different microarray technologies or normalization procedures involved, we introduced an alternative computation method described in the Supplementary Information of [Desmedt et al., 2008]. This method consists in using the robust model presented in Section 4.1.3, i.e. a signed average of the expressions of the genes included in the signature. This makes it possible to compute all the gene signatures in all the datasets in order to perform a thorough performance assessment and comparison analysis.

Each gene module score and prognostic gene signature was rescaled to derive comparable scales between datasets, as in Section 5.3. The scores were scaled such that quantiles 2.5% and 97.5% equaled to -1 and +1 respectively. This scaling was robust to outliers and ensured the scores lay approximately in $[-1, +1]$.

5.4.1.3 Results

Gene signatures We used the same gene modules as those identified in Section 5.3. They represent key biological processes in breast cancer such as proliferation, tumor invasion, immune response, angiogenesis, apoptosis, and estrogen and HER2 signaling pathways (Figure 5.10).

Additionally, we computed several existing prognostic gene signatures, namely GENE70 [van't Veer et al., 2002], GENE76 [Wang et al., 2005], P53 [Miller et al., 2005], WOUND [Chang et al., 2004], GGI [Sotiriou et al., 2006], ONCOTYPE [Paik et al., 2004] and IGS [Liu et al., 2007]

Performance assessment Like for the experimental findings of the GGI study (Section 5.2.1), we present here updated experimental findings using different datasets and a performance criterion different from the one used in the original publication [Desmedt et al., 2008].

The updated results were generated using recent datasets of untreated early (node-negative) breast cancer patients, as we did for the GGI study (Section 5.2.1) and the identification of breast cancer molecular subtypes (paragraph on clinical outcome in Section 5.3), i.e. 745 patients selected from the NKI, TBG, UPP, UNT and MAINZ datasets (Table 5.1). In the original article [Desmedt et al., 2008], untreated node-positive patients were considered as well, but the population of patients under study at that time was affected by breast cancer at different stages (heterogenous number of lymph nodes involved). This prevented us from focussing on early breast cancer, which is more interesting from a clinical point of view. For the new study, we considered DMFS as the survival endpoint.

The hazard ratio was used in the original article [Desmedt et al., 2008], but in this thesis, we highlight the drawbacks of using hazard ratio as a performance criterion (Section 4.4.1). Since the concordance index has desirable properties compared to hazard ratio (Section 4.4.1), in the study described here below we used it to assess the performance of the clinical variables, the gene modules and the state-of-the-art prognostic gene signatures.

Clinical variables and gene modules We assessed the performance of the clinical variables and the gene modules. All the clinical variables were discrete (age at diagnosis < 50 years, tumor size < 2 cm, ER-negative/positive and histological grade 1, 2 and 3). The gene modules and the signatures (see next section) were analyzed as continuous values (risk score predictions). Note that although the scales of the concordance indices were similar for continuous and discrete variables, the discretization of a variable might lead to higher concordance indices with wider confidence intervals [Harrell et al., 1996]. One has to keep this in mind when comparing discrete variables (e.g. clinical variables) with continuous variables (e.g. gene modules or gene signatures).

The forestplot represented by Figure 5.16 corresponds to Figure 3 in [Desmedt et al., 2008]. This allows us to compare the prognostic values of the clinical variables and the gene module scores with respect to the breast cancer molecular subtypes. The Table B.2 in Appendix B.2.1 reports the concordance indices, their confidence intervals and their significances. In the global population, as expected, we observed highly significant concordance indices for all the clinical variables, especially for histological grade. For the gene modules, the proliferation module (AURKA) yielded a high concordance index with a small confidence interval, confirming the importance of measuring proliferation at the molecular level,

as emphasized in [Sotiriou et al., 2006]. In the ER+/HER2- subtype, including 69% of the patients (Figure 5.15), this trend was even stronger. This result suggested that the search for prognostic factors in the global population of patients is mainly driven by the tumors of ER+/HER2- subtype, hiding potentially interesting prognostic factors in the ER-/HER2- and HER2+ subtypes. Indeed, in these subtypes, the picture dramatically changes. In the ER-/HER2- subtype, only the immune response gene module (STAT1) yielded a significant concordance index. A recent study also showed the prognostic value of the immune response quantified by gene expression profiling in ER-negative patients [Teschendorff et al., 2007; Teschendorff and Caldas, 2008]. In fact, ER status was also significant, but its prognostic value is questionable since very few patients are ER-positive in this molecular subtype. In the HER2+ subtype, only the immune response and the angiogenesis gene modules (STAT1 and VEGF respectively) yielded significant concordance indices. We did not observe a significant performance for the tumor invasion gene module (PLAU) in this subtype, in contrast to the original article [Desmedt et al., 2008], suggesting that this factor is not prognostic in early (node-negative) breast cancer.

Prognostic gene signatures We assessed here the performance of existing prognostic gene signatures. The concordance indices are represented in Figure 5.17. The Table B.3 in Appendix B.2.1 reports the concordance indices, their confidence intervals and their significances. This table corresponds to Table 2 in [Desmedt et al., 2008]. In the global population and in the ER+/HER2- subtype, all the gene signatures yield significant concordance indices. However, none is significant in the ER-/HER2- and HER2+ subtypes, suggesting the prognostic value of these gene signatures are limited to the ER+/HER2- subtype. Actually, we have shown in [Wirapati et al., 2008] that many proliferation-related genes are included in these signatures and that these genes recapitulate their prognostic value (see Figure 3 in [Wirapati et al., 2008]).

5.4.1.4 Findings

In order to reveal the thread connecting molecular subtypes, prognostic gene signatures, and traditional clinico-pathological prognostic factors, we introduced the concept of gene modules associated with key biological processes in breast cancer tumorigenesis. Wishing to extend our previous results on the GGI signature [Sotiriou et al., 2006], capturing mainly proliferation, we built several other gene modules representing key biological processes in breast cancer such as proliferation, tumor invasion, immune response, angiogenesis, apoptosis, and ER and HER2 signaling pathways.

We recapitulate below the main findings of this study:

- We showed that the gene modules we developed contain distinct prognostic information according to different breast cancer molecular subtypes, and we highlighted the importance of proliferation-related genes in predicting clinical outcome. Here we detail the findings related to the gene modules with respect to the subtypes:
  - In the ER+/HER2- subtype, the proliferation gene module and histological grade were the two most significant prognostic factors. This is consistent with our finding that two clinically distinct ER-positive molecular subtypes can be defined by the GGI, which captures mainly proliferation [Loi et al., 2007].

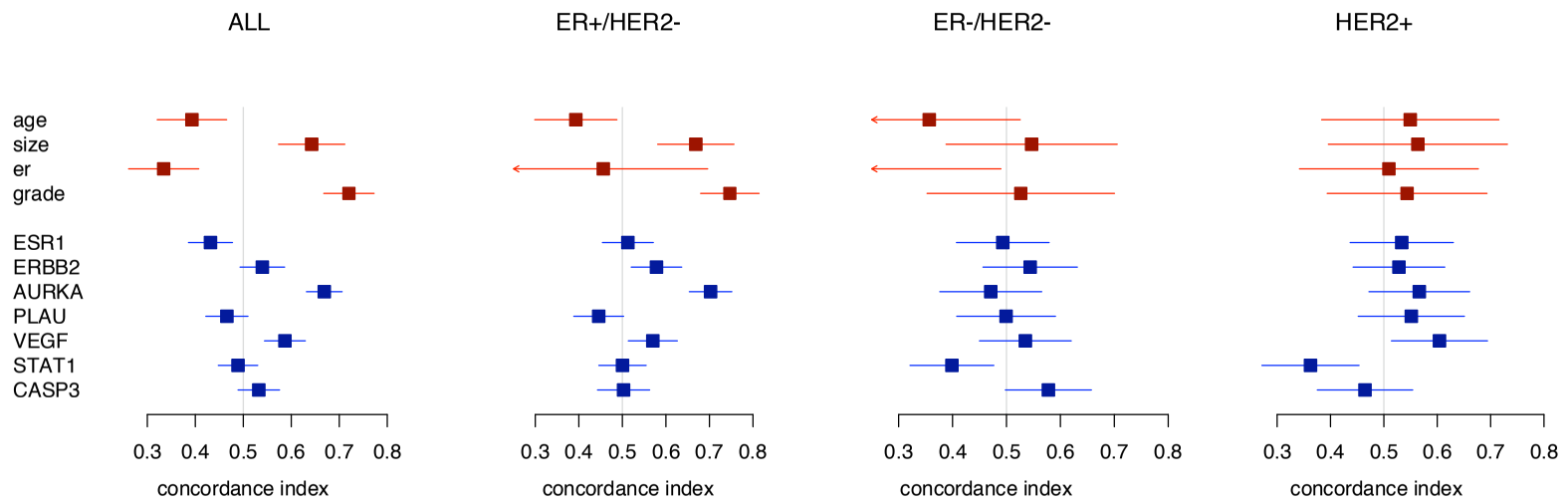


Figure 5.16: Forestplot of the concordance indices of the clinical variables (in red) and the gene module scores (in blue) with respect to the breast cancer molecular subtypes.

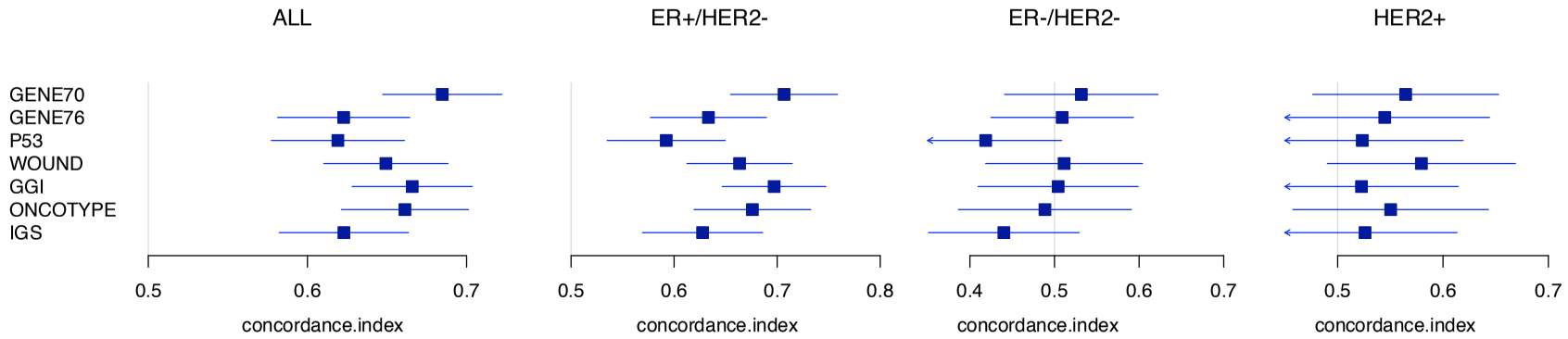


Figure 5.17: Forestplot of the concordance indices of the gene signatures with respect to the breast cancer molecular subtypes. GENE70: [van't Veer et al., 2002]; GENE76: [Wang et al., 2005]; P53: [Miller et al., 2005]; WOUND: [Chang et al., 2004]; GGI: [Sotiriou et al., 2006]; ONCOTYPE: [Paik et al., 2004]; IGS: [Liu et al., 2007].

  - In the ER-/HER2- subtype, only immune response appeared to be prognostic. It has been reported that tumors that do not express the ESR1 and ERBB2 genes, also called "basal-like" tumors, are more aggressive [Perou et al., 2000; Sorlie et al., 2001, 2003; Sotiriou et al., 2003]. It is worth mentioning that patients with basal-like tumors cannot be treated with the conventional targeted therapies currently available for breast cancer, such as endocrine or ERBB2 therapies, leaving chemotherapy as the only option. In this study, we showed that in this subtype, impaired immune response might be linked with the development of distant metastases. Indeed, high expression levels of the immune response module were associated with a significantly better outcome. Interestingly, Teschendorff et al. recently published similar findings [Teschendorff et al., 2007; Teschendorff and Caldas, 2008].
  - In the HER2+ subtype, immune response and angiogenesis appeared to be the main processes associated with breast cancer prognosis.

- Our study also highlighted that proliferation-related genes are the main and common denominator of several previously published gene signatures for predicting clinical outcome. Since defects in cell cycle deregulation are a fundamental characteristic of breast cancer, it is not surprising that these genes are involved in breast cancer prognosis. Several studies have indeed showed that increased expression of cell-cycle and proliferation-associated genes was correlated with poor outcome (reviewed in [Colozza et al., 2005]). There are of course differences in the exact proliferation-associated genes, due to the difference in population analyzed or platform used. Although the use of proliferation-associated cell markers is not new – the protein expression levels of Ki67 and PCNA have already been used as prognostic markers for decades – gene expression profiling studies suggest that measuring proliferation with a more objective, automated and quantitative assay may be more robust than less quantitative assays such as immunohistochemistry.

- We have also showed that the prognostic ability of several prognostic gene signatures differ according to the breast cancer subtypes. Indeed, we showed that their prognostic discriminative power was limited essentially to the ER+/HER2- subtype. The fact that the prognostic factors depend on the molecular subtypes highlights the importance of integrating such a knowledge into the search for new prognostic factors. Until recently, most of the prognostic gene signatures were identified using the global population of breast cancer patients.

Regarding these findings, we think that it is time now to look for gene signatures that are prognostic in specific breast cancer subtypes, especially for the ER-/HER2- subgroup, which is associated with poor prognosis and limited therapeutic options. Therefore, we strongly believe that studying the immune response mechanisms in this particular subgroup of patients might help us to better understand these tumors and to develop efficient novel targeted therapies.

5.4.2 Gene Expression Prognostic Index Using Subtypes (GENIUS)

5.4.2.1 Motivations

We know from the early gene expression studies that breast cancer can be classified into three molecular subtypes, depending mainly on the ER and HER2 phenotypes. Using the novel subtype clustering model we developed, we are now able to accurately classify breast tumors and to estimate the probability of their belonging to each of these subtypes.

From a prognostic point of view, several global prognostic gene signatures have been identified [van't Veer et al., 2002; Sotiriou et al., 2006; Naderi et al., 2007]. Since we showed through the study of gene modules (see Section 5.4.1) that these signatures are only prognostic in the ER+/HER2- subtype and are driven by proliferation-related genes, there is still room for improvement. Additionally, few gene signatures, including our own gene modules, were shown to be prognostic in specific subtypes only [Wang et al., 2005; Teschendorff et al., 2007; Desmedt et al., 2008; Finak et al., 2008].

In 2005, Wang et al. were the first to propose the development of a prognostic model by dividing the global population of patients into subgroups based on their ER status [Wang et al., 2005]. Although the approach seemed appealing and their GENE76 signature yielded a good performance, there was still room for improvement. First, the authors considered only two subgroups of patients (ER- and ER+) without taking into account the heterogeneity of HER2+ tumors. Second, the prognostic model specifically developed for ER- tumors was trained on few samples (35) and yielded poor performance in validation studies [Foekens et al., 2006; Desmedt et al., 2007].

Given these limitations, we sought to develop a novel prognostic model that would take into account the molecular heterogeneity of breast. This model, called GENIUS (Gene Expression progNostic Index Using Subtypes), should exhibit significant improvement in the prognostication of the global population of breast cancer patients and yield good performance in each molecular subtype.

The experimental findings we present below have been submitted for publication in [Haibe-Kains et al., 2009].

5.4.2.2 Methods

In order to build the new prognostic model taking into account the molecular heterogeneity of breast cancer, we followed the modular modeling approach described in Section 4.3.1. Therefore, we had to define the modules and the local models.

The modules were defined as the breast cancer molecular subtypes. The local basis functions ρ_j of Equation (4.13) were then defined as the functions returning the posterior probability of a patient to have a tumor of subtype *j* given the genetic profile of the tumor as in Equation (4.14). To estimate this posterior probability, we used the subtype clustering model developed in Section 5.3.

We built the local models by adapting the method developed for the global prognostic gene signatures (Section 4.1):

1. Feature transformation: The original gene expressions were used for the following steps of signature extraction, no feature transformation being performed.

2. Feature selection: The feature ranking procedure described in Algorithm 4 was used with a scoring function based on the weighted concordance index (Section 4.3.3.1). The weights were defined as the probabilities to belong to the subtype under study. Once the feature ranking was performed, the signature size was selected based on the stability criterion $Stab_{adj}$ (Section 4.1.2.1). We refer hereafter to these signatures as the *subtype signatures*.

3. Model building: The local models were built from the signatures as the combination of univariate models described in Section 4.1.3.1. We refer hereafter to the risk score predictions of the local models as the *subtype risk scores*.

We focussed our survival analysis on the DMFS of untreated node-negative patients in order to build a prognostic model for early stage breast cancer and to avoid any confounding factors due to the treatment effects on survival.

5.4.2.3 Results

We used VDX (Table 5.1) as the training set, since this population contains the largest sets of ER-/HER2- (99), HER2+ (54) and ER+/HER2- (191) tumors from untreated node-negative patients.

Many prognostic gene signatures have already been published on the basis of the global breast cancer population, and we have shown in Section 5.4.1 that these signatures have only added information in the ER+/HER2- subtype, and that proliferation-related genes are their common denominator. Given the considerable level of prognostic evidence in this subtype, we do not generate a new prognostic gene signature for these ER+/HER2- tumors, but consider instead the proliferation gene module (AURKA) as subtype signature.

On the contrary, very few prognostic signatures have been reported thus far in the ER-/HER2- and HER2+ subtypes [Wang et al., 2005; Teschendorff et al., 2007; Desmedt et al., 2008; Finak et al., 2008]. Therefore, we developed subtype signatures for ER-/HER2- and HER2+ tumors.

Figure 5.18 shows the design of GENIUS, inspired from the design of modular modeling in Figure 4.9. We can see a representation of the three breast cancer molecular subtypes identified by our subtype clustering model and the use of the proliferation gene module (AURKA) as prognostic signature for the ER+/HER2- subtype.

We identified the prognostic genes for the ER-/HER2- and HER2+ subtypes separately, the most stable signatures including 63 and 22 genes respectively (Figure 5.19). The genes selected for each subtype signature are given in Appendix C.4.

In order to gain biological insight into the subtype signatures used in GENIUS, we analyzed the three lists of genes using the ingenuity pathway analysis [Ingenuity Systems]. Genes from the ER-/HER2+ signature were significantly associated with functions such as cell proliferation and death, cellular movement, molecular transport, immune response and cell-to-cell interactions. Genes included in the HER2+ subtype signature were significantly associated with cellular growth and proliferation, immune response and cell signaling. The AURKA module, which was used as ER+/HER2- subtype signature, represents mainly cell cycle and proliferation genes, as reported previously [Desmedt et al., 2008].

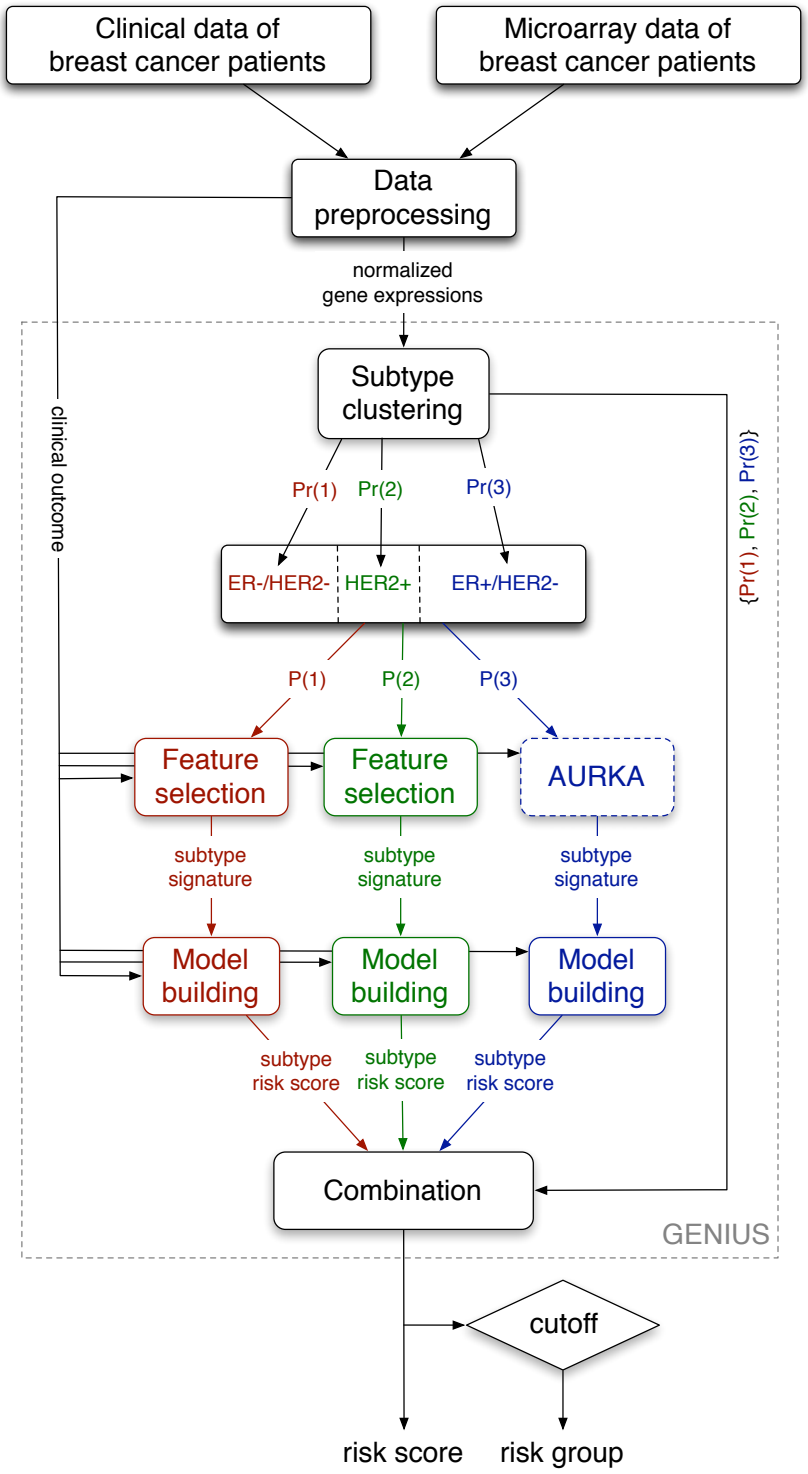


Figure 5.18: Design of GENIUS (Gene Expression progNostic Index Using Suvtypes) for breast cancer prognostication. The figure is adapted from Figure 4.9.

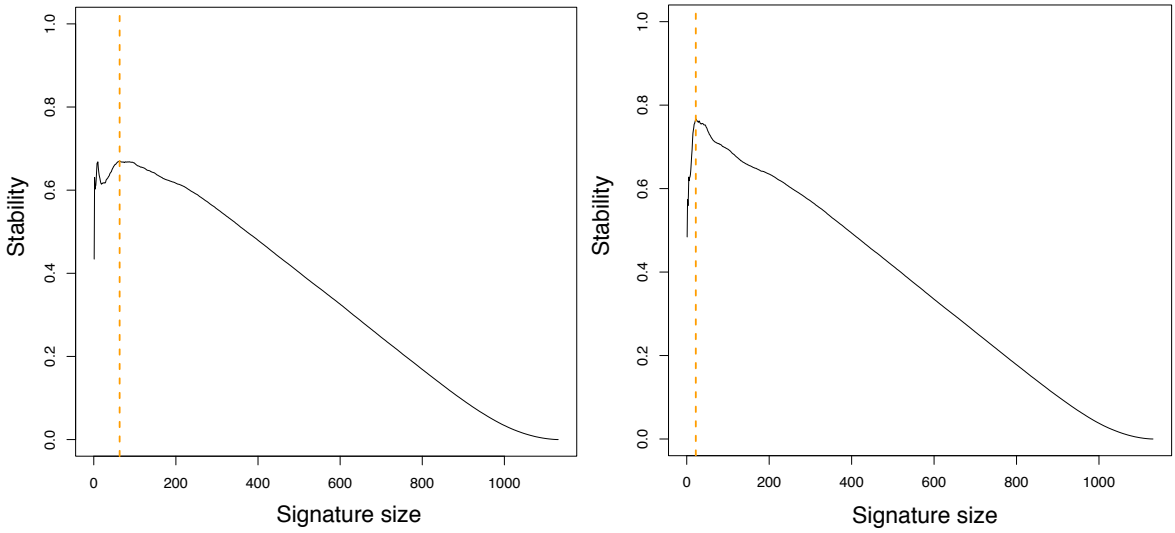


Figure 5.19: Evolution of the stability criterion $Stab_{adj}$ with respect to size in (a) the ER-/HER2+ subtype and (b) the HER2+ subtype. The vertical orange dashed lines represent the signature size maximizing the stability.

Performance assessment and comparison We assessed the performance of GENIUS in our validation set, which includes 745 node-negative untreated patients from the NKI, TBG, UPP, UNT and MAINZ datasets (Table 5.1).

Furthermore, in order to assess whether GENIUS would add prognostic information to the one provided by already published gene signatures, we compared its performance with several signatures shown to be associated with prognosis in the global breast cancer population or in a specific molecular subtype: (i) the GGI representing the initially published prognostic signatures for the global population of breast cancer patients (e.g. GENE70 or GENE76) since we showed that they all yield similar performances [Haibe-Kains et al., 2008b]; (ii) IRMODULE identified by Teschendorff et al. in the ER-negative breast cancers [Teschendorff et al., 2007; Teschendorff and Caldas, 2008]; (iii) SDPP representing the stroma-derived prognostic predictor identified by Finak et al. and shown to perform well in ER+ and HER2+ tumors [Finak et al., 2008]; (iv) the in-silico derived AURKA, PLAU and STAT1 gene modules, since our group showed that the proliferation gene module AURKA was prognostic in the ER+/HER2- subtype only and that the immune response gene module (STAT1) was prognostic in the ER-/HER2- and HER2+ subtypes, while the tumor invasion gene module (PLAU) was prognostic in the HER2+ subtype only[4] [Desmedt et al., 2008].

We compared the performance of these signatures with GENIUS in the global population and in the three molecular subtypes in our validation set. To avoid the risk of overfitting, we did not consider NKI in our validation set for IRMODULE, since this signature was identified using the NKI dataset

[4]The updated results presented in Section 5.4.1 suggested that in a population of patients with early (node-negative) breast cancers, the tumor invasion gene module (PLAU) is not prognostic. Since this result was not reported in previous publications, we included this gene module in the current study.

In the following sections, we will assess the performance for the risk score predictions only, leaving aside the risk group predictions. However, the performance assessment and comparison for the risk group predictions are available in [Haibe-Kains et al., 2009].

GENIUS vs state-of-the-art prognostic gene signatures In order to compare our risk prediction model with other gene signatures shown to be prognostic in specific breast cancer molecular subtypes, we computed the risk predictions of these signatures using the alternative computational method introduced in [Desmedt et al., 2008]. Although this method may differ from the algorithms used in the original publications, it is able to compute risk score and risk group predictions in datasets using different microarray platforms and normalizations while yielding similar performance [Desmedt et al., 2008].

In this section, we consider the risk score predictions of GENIUS and the published prognostic gene signatures. Figure 5.20 demonstrates the performance of GENIUS and the gene signatures. The concordance indices, their confidence intervals and their significances are reported in Table B.4 (Appendix B.2.2.1).

First, we observed that GENIUS is significantly associated with prognosis in the global breast cancer population, as well as in each molecular subtype. In the global population, GENIUS yielded a concordance index of 0.71. In the ER+/HER2-, ER-/HER2- and HER2+ subtypes, GENIUS yielded a concordance index of 0.70, 0.66 and 0.66 respectively (all p-values < 0.001).

Next, we compared the prognostic performance of GENIUS to the current gene signatures (Figure 5.20). GENIUS exhibited significantly better performance in the global population of patients compared to all the evaluated gene signatures. However, depending on the signature, the superiority of GENIUS was not always significant in the subtypes in which the individual signatures were originally shown to be prognostic. For example, STAT1 and IRMODULE were highly prognostic in the ER-/HER2- and HER2+ subtypes, while SDPP was associated with prognosis in the ER+/HER2- and HER2+ subtypes.

In order to explain these findings, we investigated the correlations between GENIUS and the other signatures. In the global population of patients, we observed a substantial correlation between GENIUS and AURKA (0.65), GGI (0.6) and SDPP (0.5). As expected, we observed for some signatures higher correlations in the subtypes in which the signatures had been shown to be prognostic: 0.53 and 0.75 for STAT1 in the ER-/HER2- and HER2+ subtypes respectively; 0.4 for PLAU in the HER2+ subtype; 0.51 and 0.72 for IRMODULE in the ER-/HER2- and HER2+ subtypes respectively; 0.89 in the ER+/HER2- subtype for GGI; and 0.46 in the HER2+ subtype for SDPP, not to mention the almost perfect correlation (0.99) between GENIUS and AURKA in the ER+/HER2- subtype, which can be explained by the fact that AURKA was used as a subtype signature for the ER+/HER2- tumors.

GENIUS vs prognostic clinical models In order to compare GENIUS with the best current prognostic clinical models, we computed the risk predictions using the Nottingham Prognostic Index (NPI; [Todd et al., 1987]) and Adjuvant! Online (AOL) version 8.0 [Ravdin et al., 2001].

GENIUS risk scores are poorly correlated to AOL and NPI risk scores, with a respective correlation of 0.27 and 0.39 in the global population. The correlations are even lower within the ER-/HER2- and HER2+ subtypes. We computed the concordance indices of AOL and NPI risk score predictions (Table B.5 in Appendix B.2.2.2) and compared them to GENIUS

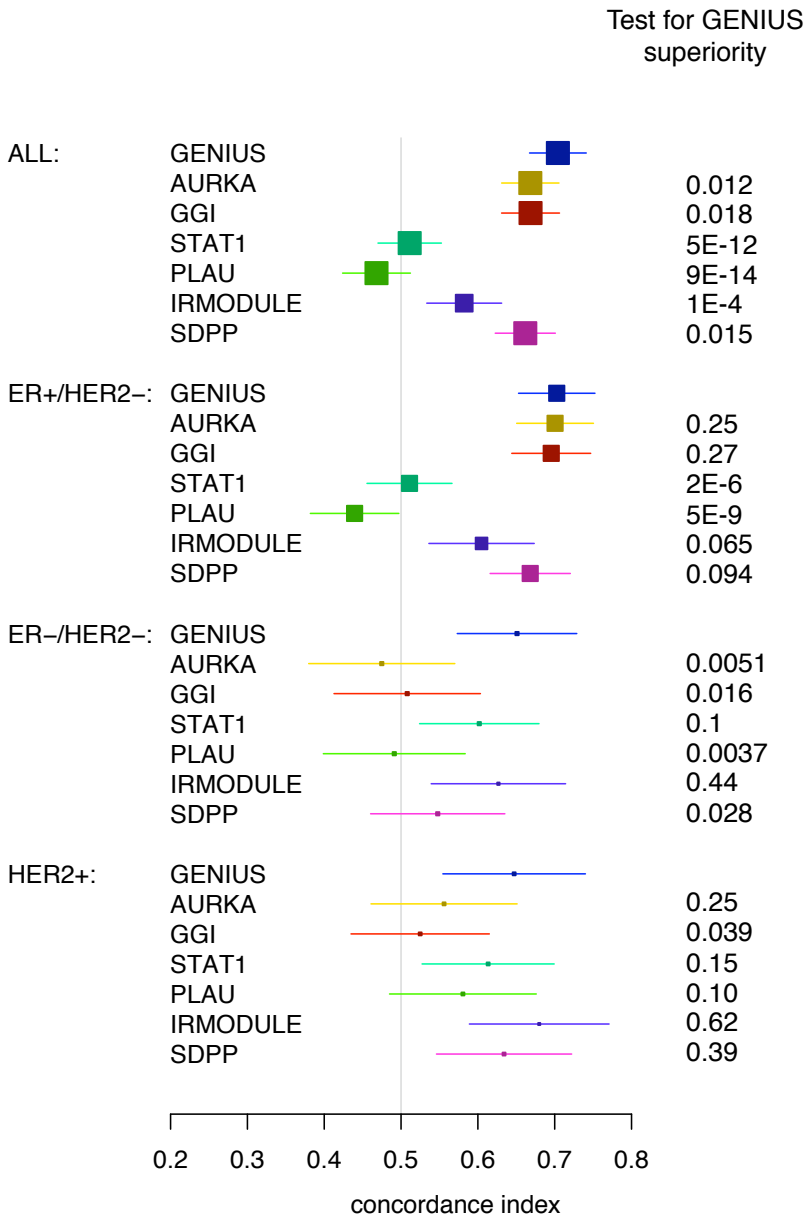


Figure 5.20: Forestplot of the concordance indices of GENIUS and the existing prognostic gene signatures with respect to the breast cancer molecular subtypes. AURKA: [Desmedt et al., 2008]; GGI: [Sotiriou et al., 2006]; STAT1: [Desmedt et al., 2008]; PLAU: [Desmedt et al., 2008]; IRMODULE: [Teschendorff et al., 2007]; SDPP: [Finak et al., 2008].

as shown in Figure 5.21. Although we observed better performance for GENIUS in the global population, its superiority did not reach significance in all molecular subtypes. NPI yielded a performance similar to GENIUS in the ER+/HER2- subtype. GENIUS was significantly better than both AOL and NPI in the ER-/HER2- subtype, while we observed only a trend for the superiority of GENIUS in the HER2+ subtype (p-value < 0.10).

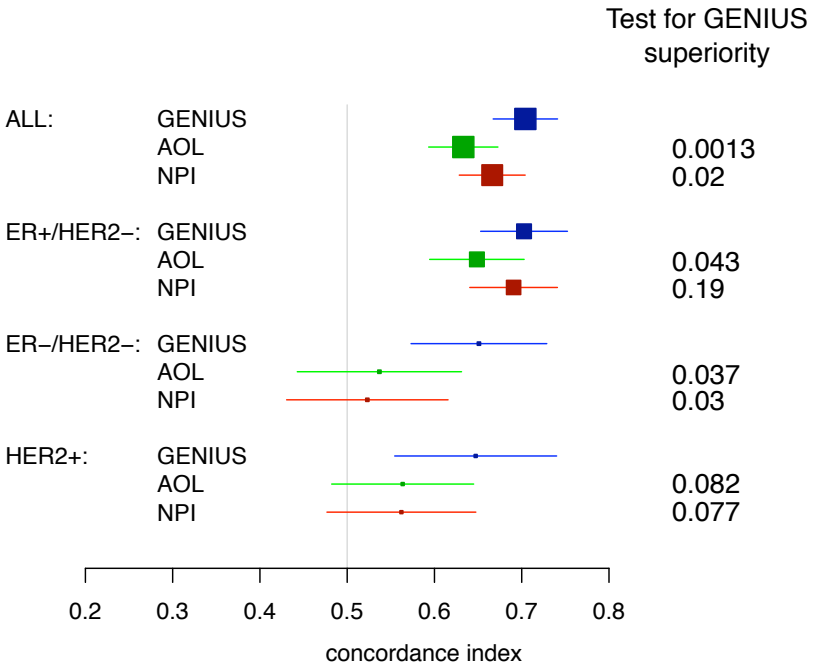


Figure 5.21: Forestplot of the concordance indices of GENIUS and the prognostic clinical models with respect to the breast cancer molecular subtypes. AOL: [Ravdin et al., 2001]; NPI: [Todd et al., 1987].

5.4.2.4 Findings

In this study, we introduced a new methodology for improving breast cancer prognostication using microarray data, by taking into account the molecular heterogeneity of breast cancer. This new risk prediction model was developed to answer the major criticism raised regarding the great majority of gene signatures reported so far, i.e. that these are only prognostic for ER-positive disease [Desmedt et al., 2008; Wirapati et al., 2008]. While it is clear that the HER2+ and ER-/HER2- breast cancer molecular subtypes of tumors have an overall worse prognosis than the ER+/HER2- ones, some of these patients do have a better clinical outcome. However, only few studies have so far attempted to consider the molecular heterogeneity of the HER2+ and ER-/HER2- breast cancer and to derive a prognostic gene signature for these breast cancer subtypes [Wang et al., 2005; Teschendorff et al., 2007; Desmedt et al., 2008].

In particular, we showed that:

- GENIUS was highly prognostic in the global population and in all breast cancer subtypes. GENIUS yielded a significantly better performance than all the state-of-the-art

prognostic gene signatures in the global population. Although GENIUS was not significantly better in all the specific subtypes – the immune response signatures, namely STAT1 [Desmedt et al., 2008] and IRMODULE [Teschendorff et al., 2007], being particularly good performers in the HER2+ subtype – it was the only prognostic gene signature to yield good performance whatever the molecular subtype.

- GENIUS yielded better performance than traditional prognostic models. Indeed, a criticism raised in recent years regarding the existing prognostic gene signatures is that they may add only little information beyond that provided by the traditional clinical guidelines. To that end, we considered the Nottingham Prognostic Index (NPI) and Adjuvant Online (AOL) as the reference for assessing the risk of recurrence using the traditional clinico-pathologic parameters. We compared these clinical guidelines and the gene expression index. The prognostic information provided by AOL and NPI seemed to be limited to the ER+/HER2- subtype. We showed that GENIUS yielded significantly better performance in the global population of breast cancer patients, improving upon the traditional clinical models.

Although we showed that GENIUS outperformed prognostic clinical models, namely AOL and NPI, we lacked clinical information in the training set (VDX) to study the complementarity of the clinical and microarray data. Indeed, several authors showed recently that the performance of prognostic gene signatures could be improved by combining them with clinical variables [Gevaert et al., 2006; Boulesteix et al., 2008; Wirapati et al., 2008; Sotiriou and Pusztai, 2009]. However, building a risk prediction model that combines clinical and microarray data is a difficult task. This is mainly due to the high dimensionality of the microarray data, which often leads to an underestimation of the relevance of the clinical variables. Gevaert et al. used an ingenious approach based on Bayesian networks to treat clinical and microarray data on an equal footing [Gevaert et al., 2006]. Were the clinical information for the VDX dataset available, such an approach could be used to combine GENIUS with clinical variables to further improve the performance of the model.

Chapter 6

Conclusions

In the early 1990s the era of high throughput technologies started, changing our way of studying biology. Using these technologies, scientists are now able to draw a global picture of the state of cells at the molecular level. In this thesis, gene expression profiling, through microarray technology, was used to study cancer cells in order to bring new insights into breast tumorigenesis and to build new predictive tools for breast cancer management.

This concluding chapter will summarize the major methodological guidelines and experimental findings made in this thesis, while emphasizing the main motivations that have driven our research.

6.1 Methodological Guidelines

The intrinsic complexity of microarray data, especially their huge dimensionality, have raised new challenges in data analysis. The development and the application of novel machine learning methods to extract knowledge from microarray data is an active research field.

In this thesis, the approach we proposed to address microarray data analysis for breast cancer prognostication was inspired by 5 main principles: (i) accuracy; (ii) biological interpretability; (iii) robustness; (iv) thorough performance assessment and comparison; and (v) research reproducibility.

Accuracy Predictive models should yield good performance in independent data. In early publications in microarray data analysis, simple to complex machine learning techniques were used. As the field became increasingly mature, large scale comparative studies showed that simple methods outperformed complex ones due to the intrinsic complexity of microarray data. However, these comparative studies focused on class discovery and classification, leaving aside methods for survival analysis, particularly suited to breast cancer prognostication. To fill this important gap, we performed a large-scale comparison study of risk prediction models and, similar to what was found for class discovery and classification from microarray data, we observed that simple methods outperformed complex ones, suggesting that variance is the most important term to reduce in the bias-variance trade-off (Section 5.2.2). Therefore we developed simple yet robust risk prediction models, taking care to reduce the variance of the models (Section 4.1.3). The rationale is that the gain in stability largely compensates for the lack of complexity.

Biological interpretability Predictive models should be interpretable from a biological point of view in order to potentially gain new biological insights into tumorigenesis. Since the aim of developing new predictive tools for breast cancer prognostication from microarray data is twofold, i.e. to yield good prediction performance and to gain new biological insights into cancer biology, we developed machine learning techniques in order to keep the final predictive model interpretable. This was made possible by properly designing the different steps of our microarray data analysis methodology. At the feature transformation step, we designed two methods to reduce the dimensionality of the data while keeping the features interpretable (Sections 4.1.1 and 4.2.1). At the feature selection step, we assessed the stability of the selection in order to reinforce the confidence of doctors in the signature identified (Section 4.1.2). At the model building step, both for global and local risk prediction models, we avoided the use of "black box" models (complex models whose interpretation is difficult) to focus on simple, yet robust predictive models (Section 4.1.3). Lastly, our local risk prediction model GENIUS, beyond providing accurate risk prediction information for doctors, also provides information about the probability that a patient belongs to a particular breast cancer molecular subtype (Sections 4.3 and 5.4.2). This information is crucial for doctors since patients with different subtypes of breast cancer will respond differently to various therapies.

Robustness The methods and the resulting predictive models should be usable in datasets involving different microarray technologies and normalization techniques. Emphasis was placed on predictive models useable in the numerous datasets we collected during the thesis (Section 5.1). Indeed, the use of the predictive models should not be limited to a specific microarray technology nor to data normalization. Instead, the model should be useable and yield good prediction performance whatever the datasets. This facilitates validation of the model since it can be developed on one dataset and validated on one or several other datasets, irrespective of the microarray platform. In this thesis, we facilitate such a validation by using feature transformation and/or model building methods. In the feature transformation step, the computation of features, which is an average of highly correlated gene expressions, facilitates the computation of such features in other microarray platforms in which some genes are not represented (Sections 4.1.1 and 4.2.1). In the model building step, the robust signed average used to predict patient risk makes possible the computation of such risk predictions in datasets using different microarray platforms (Section 4.1.3).

Thorough performance assessment and comparison In order to assess the performance of our new risk prediction models and compare them with the state-of-the-art, we developed in this thesis a statistical framework for performance assessment and comparison (Section 4.4). We implemented in a tool all the functions required to assess and compare the performance risk prediction models, such as the prognostic clinical models and the prognostic gene signatures (Section 4.4). A thorough performance assessment and comparison of risk prediction models is particularly important in the field of breast cancer prognostication using microarray data, in which the improvement brought by a gene signature has to be large enough to compensate for the cost of microarray experiments and the complexity related to their use in day-to-day clinical practice.

Research reproducibility We also emphasized throughout the thesis the importance of the research reproducibility guidelines proposed in [Gentleman, 2005]. Indeed, microarray

data analysis involves many steps and even a careful analyst may make mistakes that can have a large impact on the final results of a study. Currently, most studies are hardly reproducible, even with the help of the authors [Ioannidis, 2005; Dupuy and Simon, 2007]. In this thesis, we provided for [Haibe-Kains et al., 2008a,c] the SWEAVE code in order to ensure the reproducibility of the results at each step of the analysis.

The application of the novel methods we developed led to important findings related to breast cancer biology and prognostication, as presented in the section below.

6.2 Experimental Findings

In this thesis, we identified both global and local prognostic gene signatures. These signatures will be briefly summarized here, while their application in clinic will be emphasized in the section dedicated to translational research.

6.2.1 Global Prognostic Gene Signatures

Global prognostic gene signatures are signatures that do not take into account the presence of breast cancer molecular subtypes.

GGI The GGI signature, presented in Section 5.2.1, is predictive of the histological grade of breast tumors [Sotiriou et al., 2006]. We showed its strong prognostic value in several independent microarray datasets. Moreover, we extensively compared its performance to numerous risk prediction models [Haibe-Kains et al., 2008c] and to two state-of-the-art prognostic gene signatures, namely GENE70 and GENE76, in a dedicated validation dataset [Haibe-Kains et al., 2008b].

From a biological point of view, the vast majority of the genes included in the GGI signature are related to proliferation.

TAMR13 The TAMR13 signature, presented in Section 5.2.3, is predictive of the resistance to tamoxifen [Loi et al., 2008]. We showed the predictive ability of this signature in three independent datasets of tamoxifen treated breast cancer patients.

From a biological point of view, the (clusters of) genes included in the TAMR13 signature are related to proliferation, tumor invasion, immune response and cellular inflammatory response.

6.2.2 Local Prognostic Gene Signatures

The local prognostic gene signatures are signatures extracted or evaluated by taking into account the presence of breast cancer molecular subtypes.

Gene modules Using *a priori* biological knowledge in the form of the selection of seven key biological processes in breast tumorigenesis, a gene signature (gene module, see Section 5.4.1) was identified for each of these processes in order to elucidate the prognostic factors with respect to the breast cancer molecular subtypes [Wirapati et al., 2008; Desmedt

et al., 2008]. We found that each molecular subtype exhibited different prognostic factors. Additionally, we showed that most state-of-the-art prognostic gene signatures (e.g. GENE70 or GENE76) yield good performance in the ER+/HER2- subtype only and that their prognostic ability is driven by the proliferation-related genes.

From a biological point of view, each gene module was well defined since their identification was driven by the selection of seven key biological processes in breast tumorigenesis, i.e. ER (ESR1) and HER2 (ERBB2) signaling, proliferation (AURKA), tumor invasion (PLAU), angiogenesis (VEGF), immune response (STAT1) and apoptosis (CASP3).

GENIUS The GENIUS signature, presented in Section 5.4.2, was identified using a modular modeling approach adapted to breast cancer prognostication [Haibe-Kains et al., 2009]. GENIUS is composed of three subtype signatures extracted from each breast cancer molecular subtype. We showed that GENIUS significantly outperforms state-of-the-art prognostic gene signatures as well as the prognostic clinical models in the global population of patients.

From a biological point of view, each subtype signature reflects different processes. The subtype signature associated with ER+/HER2- was chosen to be the proliferation gene module. The subtype signature associated with ER-/HER2- contains genes related mainly to proliferation, immune response and cell-to-cell interactions. The subtype signature associated with HER2+ contains genes related mainly to immune response and cell signaling.

6.2.3 Biological Insights

The application of the methods developed in this thesis led to efficient predictive models, while bringing new insights into breast cancer biology.

From the identification of global gene signatures, we learned the importance of proliferation genes for prognostication. Proliferation had already been known for years to be highly prognostic. However, we showed that the robust quantification of proliferation genes through microarray technology yields better performance than traditional histo-pathological measurements (e.g. histological grade). In addition, we showed that proliferation is also relevant for the prediction of tamoxifen resistance.

From the identification of local gene signatures, we learned that the prognostic factors depend on the breast cancer molecular subtypes. Proliferation is strongly prognostic in the ER+/HER2- subtype only, immune response is prognostic in the ER-/HER2- and HER2+ subtypes, and angiogenesis is prognostic in the HER2+ subtype. We also observed that, due to the large proportion of ER+/HER2- patients and the strong prognostic value of proliferation related genes, the prognostic value of most state-of-the-art gene signatures is driven by these genes only. The local gene signatures allowed us to highlight other biological processes involved in breast cancer prognosis, depending on the molecular subtypes.

In addition to improving our understanding of breast tumorigenesis, efforts have been made to bring some of the gene signatures identified in this thesis into the clinic in order to provide real benefit to breast cancer patients. This transfer of knowledge from the laboratory to day-to-day clinical practice is called *translational research* and is presented below.

6.2.4 Translational Research

One of the objectives of the Functional Genomics Unit at Institut Jules Bordet in Brussels, headed by Prof. Christos Sotiriou is to bring the prognostic gene signatures we identified into routine clinical practice. To do so, patents for the GGI signature and the tumor invasion (PLAU) and immune response (STAT1) modules were deposited. The two patents related to the gene modules are under validation review. The patent for GGI has recently been filled (full description of the patent can be accessed from `http://www.wipo.int/pctdb/en/wo.jsp?wo=2006119593`). The bibliographic data for the GGI patent are provided below.

(WO/2006/119593) GENE-BASED ALGORITHMIC CANCER PROGNOSIS

Biblio. Data | Description | Claims | National Phase | Notices | Documents

Latest bibliographic data on file with the International Bureau

Pub. No.: WO/2006/119593 **International Application No.:** PCT/BE2006/000051
Publication Date: 16.11.2006 **International Filing Date:** 15.05.2006
Chapter 2 Demand Filed: 13.03.2007

IPC: ***G06F 19/00*** (2006.01)

Applicants: **UNIVERSITE LIBRE DE BRUXELLES** [BE/BE]; Avenue Franklin Roosevelt 50, CP 161, 1050 Brussels (BE) *(All Except US)*.
SOTIRIOU, Christos [GR/BE]; (BE) *(US Only)*.
DELORENZI, Mauro [CH/CH]; (CH) *(US Only)*.
PICCART, Martine [BE/BE]; (BE) *(US Only)*.

Inventors: **SOTIRIOU, Christos**; (BE).
DELORENZI, Mauro; (CH).
PICCART, Martine; (BE).

Agent: **VAN MALDEREN, Joëlle**; pronovem-OFFICE VAN MALDEREN, Avenue Josse Goffin 158, B-1082 Brussels (BE).

Priority Data: 60/680,543 13.05.2005 US
05447274.1 07.12.2005 EP

Title: GENE-BASED ALGORITHMIC CANCER PROGNOSIS

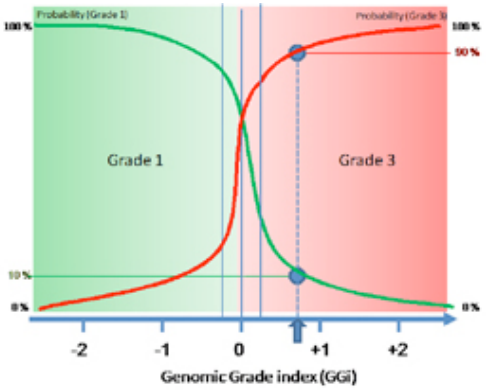
Abstract: The present invention is related to The methods and systems for prognosis determination in tumor samples, by measuring gene expression in a tumor sample and applying a gene-expression grade index (GGI) or a relapse score (RS) to yield a numerical risk score.

The GGI patent was sold through the ”ULB-Interface” and is currently commercialized by a French biotechnology company, namely IPSOGEN[1], to aid in treatment decision-making for early breast cancer patients with histological grade 2 tumors (Figure 6.1).

[1] `http://www.ipsogen.com/breast-cancer-products/healthcare-professionnals/mapquant-dxgenomicgrade/`

Improving the clinical value of tumor grading

MapQuant Dx™ Genomic Grade is the cornerstone of the MapQuant Dx™ assay series. It is the very first, microarray-based and clinically-validated, molecular diagnostic test to accurately measure tumor grade, a consensus indicator of tumor proliferation, risk of metastasis and response to chemotherapy.



Resolving histological grading uncertainty

Tumor grade is a decision factor in most national & international guidelines to breast cancer treatment. It is generally recommended to treat high-grade "grade 3" breast carcinoma with chemotherapy because they are chemosensitive and will often recur otherwise. By contrast, most low-grade "grade 1" tumors should not be treated with chemotherapy because they have a good prognosis and often are chemo-insensitive.

A critical clinical issue is how to treat the 50% of breast cancers tested today as intermediate grade?

The MapQuant Dx™ Genomic Grade test now allows to resolve more than 80% of these uncertain "grade 2" tumors into "grade 1" or "grade 3" tumors, potentially sparing useless chemotherapy treatments to tens of thousands patients a year.

Clinical utility of the Genomic Grade index

Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis.
J Natl Cancer Inst. 2006 Feb 15;98(4):262-72.
Sotiriou C,Wirapati P, Loi S, Harris A, Fox S, Smeds J, Nordgren H, Farmer P, Praz V,Haibe-Kains B, Desmedt C, Larsimont D,Cardoso F, Peterse H, Nuyten D, Buyse M, VandeVijver MJ , Bergh J , Piccart M , Delorenzi M.

Definition of clinically distinct molecular subtypes in estrogen receptor-positive breast carcinomas through genomic grade.
J Clin Oncol. 2007 Apr1; 25(10):1239-46.
Loi S, Haibe-Kains B, Desmedt C, Lallemand F, Tutt AM, Gillet C, Ellis P, Harris A, Bergh J, Foekens JA, Klijn JG, Larsimont D, Buyse M, Bontempi G, Delorenzi M, Piccart MJ, Sotiriou C.

Figure 6.1: GGI commercialized by the French biotech company IPSOGEN. Screenshot of IPSOGEN website.

6.3 Future Works

This section aims to outline the future work to be done in order to improve and extend the methods and experimental findings presented in this thesis.

- R package: In addition to the `survcomp` package, another R package is in preparation. This package, called `bcclassifier`, will implement the computation procedure of numerous state-of-the-art prognostic gene signatures, as well as those developed in this thesis. It is of utmost importance for an analyst to be able to compare a new prediction model with the state-of-the-art. However, this procedure can be tedious, since the signature and the structure of the predictive model are not always detailed in the initial publications. This package aims to implement numerous state-of-the-art prognostic gene signatures such as GENE70, GENE76, ONCOTYPE, P53, WOUND, and the gene signatures identified in this thesis, i.e. GGI, TAMR13, Gene Modules, and GENIUS. Moreover, the alternative computation procedure introduced in the Supplementary Information of [Desmedt et al., 2008] will be implemented as well in order to allow the analyst to compute risk prediction, irrespective of the microarray platforms or normalization techniques used. Lastly, this package will also implement Perou's method and our subtype clustering model for the identification of breast cancer molecular subtypes.

- Interface for `R` packages: Although the implementation of methods in R packages enables scientists familiar with this programming language to use them for their own research, R may be not "'ergonomic"' enough for some doctors or biologists. The integration of the R packages into existing stand-alone or web-based software suites will increase the accessibility of these tools. TM4[2] is a good candidate for such an integration since this open-source microarray software suite, implemented in JAVA, makes it possible for the analyst (doctors, biologists or bioinformaticians) to easily perform basic microarray analyses, from data normalization to unsupervised and supervised analyses.

- Other cancers: Although we focus our research on breast cancer, our methods could be applied to prognostication and prediction for other types of cancer. In particular, we plan to extend the GENIUS approach to lymphoma, the molecular subtypes of which have recently been described [Wright et al., 2003].

- Prediction in the neoadjuvant setting: We developed our method to deal with survival data. This choice was motivated by the fact that this type of data is used for prognostication and prediction in the adjuvant setting. Since increasing quantities of data for prediction in the neoadjuvant setting are publicly available (high throughput data as input variables and resistance/response to treatment as output variable), it would be interesting to test whether our methods apply successfully to classification.

- Genome-wide feature transformation: A more sophisticated integration of the biological annotations in the genome-wide feature transformation method could improve the performance and the biological interpretation of the method. Indeed, the current implementation discards clusters containing too few genes annotated in databases. Improvements could derive from the development of a method reordering the dendrogram

[2] `http://www.tm4.org/mev.html`

computed from gene expression data, based on biological annotations (e.g. gene ontology; [Ashburner et al., 2000]) or even a novel clustering algorithm taking into account the gene expression data and the biological annotations simultaneously.

- Stability-based feature selection: Feature selection methods dealing efficiently with the complementarity and redundancy of features could improve the way we currently implement stability-based feature selection. The methods based on information theory for feature selection [Meyer, 2008] could be adapted for survival analysis to address this issue.

- Robust model building: The risk prediction models we designed in this thesis facilitate the computation of risk predictions in other microarray platforms where some features are missing. This is important in microarray data analysis, where risk prediction models are often validated in datasets generated by different microarray platforms in which some genes are not represented. It would be interesting to assess the actual impact of missing features (Which and how many features are missing?) on final risk prediction performance. Moreover the use of robust statistics (e.g. trimmed average) might improve the performance of the risk prediction model in the presence of outliers in microarray data.

- Robust regression: In robust statistics [Hampel et al., 2005], robust regression is a form of regression analysis designed to deal with the presence of outliers [Andersen, 2007]. Although several robust regression methods have been introduced for traditional regression (e.g. least median squares [Rousseuw, 1984] or least trimmed squares [Rousseeuw and Leroy, 2003]), only a few have been proposed in survival analysis, especially for the Cox model [Lin and Weil, 1989; Leon et al., 2005]. The use of such a method for risk prediction as well as the comparison of its performance for breast cancer prognostication using microarray data, are of high interest.

- Prototype-based feature transformation: A limitation of the prototype-based feature transformation method is the fact that only one prototype can be selected to represent each biological process of interest. An adaptation of the *gene recommender* algorithm [Owen et al., 2003] would circumvent this difficulty. Moreover, it would be interesting to compare the performance of prototype-based clustering with other clustering algorithms, like quality-based clusterings [De Smet et al., 2002; Tseng and Wong, 2005].

- Subtype clustering model: Our method to identify breast cancer molecular subtypes uses model-based clustering in a two-dimensional space defined by the ESR1 and ERBB2 module scores. Although this model is robust, it has not allowed us to discern new breast cancer molecular subtypes. The use of a larger number of dimensions would lead to the discovery of new and robust molecular subtypes. Moreover, our subtype clustering model could be adapted to other cancers, such as lymphoma, the different molecular subtypes of which have recently been discovered [Wright et al., 2003].

- Local Model Network: The modular modeling approach we adopted for the identification of the local prognostic gene signatures might be especially interesting for the risk prediction of patients having a tumor whose subtype is not clearly identified (e.g. several posterior probabilities of subtype belonging greater than 20%). Indeed, the com-

bination of the local risk prediction models might be more informative than the model specific for the most likely subtype alone. A study could be conducted to highlight such an advantage of our approach compared with a crisp partition of the input space with respect to the maximum posterior probability of subtype belonging.

- Combination with clinical variables: Several authors showed recently that the performance of prognostic gene signatures could be improved by combining them with clinical variables [Gevaert et al., 2006; Boulesteix et al., 2008; Wirapati et al., 2008]. However, building a risk prediction model that combines clinical and microarray data is a difficult task. This is mainly due to the high dimensionality of the microarray data, which often leads to an underestimation of the relevance of the clinical variables. Gevaert et al. used an ingenious approach based on Bayesian networks to treat clinical and microarray data on an equal footing [Gevaert et al., 2006]. The use of this framework in combination with the GENIUS method could improve the current model for breast cancer prognostication. As a preliminary study, it would be interesting to build a risk prediction model by adopting a simple *majority voting* scheme [Hastie et al., 2001; Dudoit et al., 2002] combining GENIUS and tumor size in order to highlight the potential performance improvement of such a combination.
- Causal inference: To improve the interpretability of prediction models using gene expression data, statistical techniques for causal inference [Sprites et al., 2000] could be used to identify the direct and indirect causes of the phenomenon under study (e.g. risk of recurrence). This could lead to clear graphical representations of the relationships between the variables included in the prediction model (e.g. genes), and therefore facilitate the understanding of the model. The Bayesian approach used in [Gevaert et al., 2006] could be adapted to the risk prediction models developed in this thesis.

6.4 Integrative Bioinformatics

The near future of bioinformatics and biomedicine will be characterized by the need to integrate increasing numbers of sources of high dimensional data. High dimensional data will be generated by new high throughput technologies, e.g. single nucleotide polymorphism (SNP) or comparative genomic hybridization (CGH), at a continuously growing rate. Confronted with this overwhelming quantity of data, doctors will demand increasingly more effective, interpretable and robust computational techniques capable of producing exploitable information from genomic data. In order to make full use of these genomic data, novel methods to integrate the different data sources have to be developed. These methods could follow two axes, as illustrated in Figure 6.2. The horizontal axis refers to the integration of different datasets generated from the same class of technology (e.g. gene expression profiling, SNP, or CGH). Although the data are similar from a biological point of view (e.g. all the microarray platforms for gene expression profiling measure gene expressions), different platforms and normalization techniques are often used to generate the datasets. Therefore, their integration is a complex task that requires a meta-analytical approach. The vertical axis refers to the integration of data generated from different technologies. This type of integration requires extensive knowledge of each technology from both a technical and a biological point of view in order to be able to efficiently combine these different sources of information. This will be a great new challenge to face in Bioinformatics!

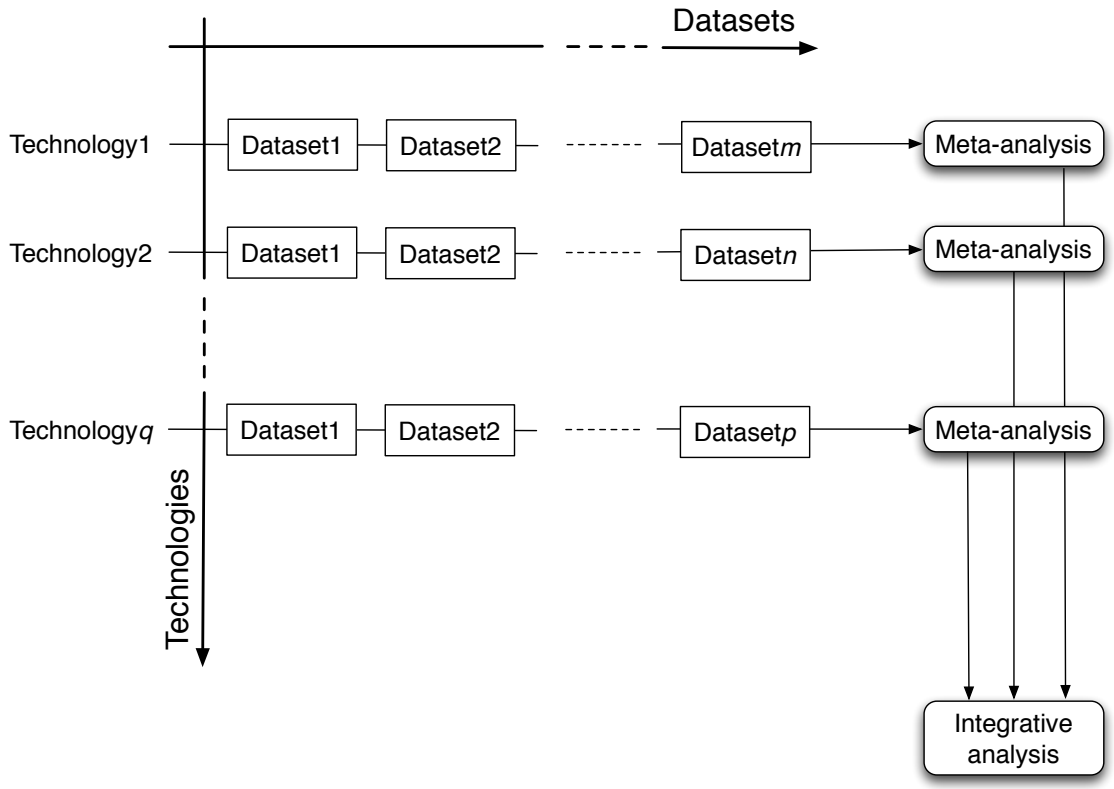


Figure 6.2: Integrative bioinformatics. The integration follows two axes: (i) Datasets: the horizontal axis refers to the integration of different datasets generated from the same class of technology (*meta-analysis*); (ii) Technologies: the vertical axis refers to the integration of data generated from different technologies (*integrative analysis*).

Bibliography

M. Adams, J. Kelley, J. Gocayne, M. Dubnick, M. Polymeropoulos, H. Xiao, C. Merril, A. Wu, B. Olde, R. Moreno, and a. et. Complementary DNA sequencing: expressed sequence tags and human genome project. *Science*, 252(5013):1651–1656, 1991. doi: 10.1126/science.2047873. URL http://www.sciencemag.org/cgi/content/abstract/252/5013/1651.

Affymetrix. Affymetrix, inc. URL http://affymetrix.com.

Affymetrix. *GeneChip Expression Analysis*, 2002.

Affymetrix. *GeneChip Expression Analysis: Data Analysis Fundamentals*, 2004. URL http://www.affymetrix.com/support/downloads/manuals/data_analysis_fundamentals_manual.pdf.

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

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

A. A. Alizadeh, M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, X. Yu, J. I. Powell, L. Yang, G. E. Marti, T. Moore, J. Hudson, L. Lu, D. B. Lewis, R. Tibshirani, G. Sherlock, W. C. Chan, T. C. Greiner, D. D. Weisenburger, J. O. Armitage, R. Warnke, R. Levy, W. Wilson, M. R. Grever, J. C. Byrd, D. Botstein, P. O. Brown, and L. M. Staudt. Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling. *Nature*, 403(6769):503–511, 2000. doi: 10.1038/35000501. URL http://dx.doi.org/10.1038/35000501.

D. M. Allen. The relationship between variable and data augmentation and a method of prediction. *Technometrics*, 16:125–127, 1974.

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

C. Ambroise and G. J. McLachlan. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proceedings of the National Academy of Sciences of the United States of America*, 99(10):6562–6566, 2002. doi: 10.1073/pnas.102102699. URL http://www.pnas.org/content/99/10/6562.abstract.

R. Andersen. *Modern Methods for Robust Regression*. Quantitative Applications in the Social Sciences. Sage Publications, Inc, 2007. ISBN 978-1412940726.

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

M. Ayers, W. Symmans, J. Stec, A. Damokosh, E. Clark, K. Hess, M. Lecocke, J. Metivier, D. Booser, N. Ibrahim, V. Valero, M. Royce, B. Arun, G. Whitman, J. Ross, N. Sneige, G. Hortobagyi, and L. Pusztai. Gene Expression Profiles Predict Complete Pathologic Response to Neoadjuvant Paclitaxel and Fluorouracil, Doxorubicin, and Cyclophosphamide Chemotherapy in Breast Cancer. *J Clin Oncol*, 22(12):2284–2293, 2004. doi: 10.1200/JCO.2004.05.166. URL http://jco.ascopubs.org/cgi/content/abstract/22/12/2284.

F. R. Bach and M. I. Jordan. Kernel independent component analysis. *J. Mach. Learn. Res.*, 3:1–48, 2003. ISSN 1533-7928.

T. Bammler, R. P. Beyer, S. Bhattacharya, G. A. Boorman, A. Boyles, B. U. Bradford, R. E. Bumgarner, P. R. Bushel, K. Chaturvedi, D. Choi, M. L. Cunningham, S. Deng, H. K. Dressman, R. D. Fannin, F. M. Farin, J. H. Freedman, R. C. Fry, A. Harper, M. C. Humble, P. Hurban, T. J. Kavanagh, W. K. Kaufmann, K. F. Kerr, L. Jing, J. A. Lapidus, M. R. Lasarev, J. Li, Y.-J. Li, E. K. Lobenhofer, X. Lu, R. L. Malek, S. Milton, S. R. Nagalla, J. P. O'malley, V. S. Palmer, P. Pattee, R. S. Paules, C. M. Perou, K. Phillips, L.-X. Qin, Y. Qiu, S. D. Quigley, M. Rodland, I. Rusyn, L. D. Samson, D. A. Schwartz, Y. Shi, J.-L. Shin, S. O. Sieber, S. Slifer, M. C. Speer, P. S. Spencer, D. I. Sproles, J. A. Swenberg, W. A. Suk, R. C. Sullivan, R. Tian, R. W. Tennant, S. A. Todd, C. J. Tucker, B. Van Houten, B. K. Weis, S. Xuan, and H. Zarbl. Standardizing global gene expression analysis between laboratories and across platforms. *Nat Methods*, 2(5):351–356, May 2005. ISSN 1548-7091 (Print). doi: 10.1038/nmeth754.

T. Barrett, T. O. Suzek, D. B. Troup, S. E. Wilhite, W.-C. Ngau, P., D. Rudnev, A. E. Lash, W. Fujibuchi, and R. Edgar. NCBI GEO: mining millions of expression profiles - database and tool. *Nucleic Acids Research*, 33:D562, 2005.

A. Ben-Dor, L. Bruhn, N. Friedman, I. Nachman, M. Schummer, and Z. Yakhini. Tissue classification with gene expression profiles. *Journal of Computational Biology*, 7(3-4): 559–583, 2000. doi: 10.1089/106652700750050943. URL http://www.liebertonline.com/doi/abs/10.1089/106652700750050943.

A. Ben-Hur, A. Elisseeff, and I. Guyon. A stability based method for discovering structure in clustered data a stability based method for discovering structure in clustered data. *Proc. Symp. Biocomput.*, 7:6–17, 2002.

A. H. Bild, G. Yao, J. T. Chang, Q. Wang, A. Potti, D. Chasse, M.-B. Joshi, D. Harpole, J. M. Lancaster, A. Berschuk, J. A. Olson Jr, J. R. Marks, H. K. Dressman, M. West, and J. R. Nevins. Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature*, 439:353–356, 2006.

M. Birattari, T. Stutzle, L. Paquete, and K. Varrentrapp. A racing algorithm for configuring metaheuristics. In W. B. Langdon, editor, *GECCO 2002*, pages 11–18. Morgan Kaufmann, 2002.

C. M. Bishop. Mixture density networks. Technical report, Astom Univeristy, February 1994.

G. Blom. *Statistical Estimates and Transformed Beta Variables*. John Wiley and Sons, 1958.

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

H. Bonnefoi, A. Potti, M. Delorenzi, L. Mauriac, M. Campone, M. Tubiana-Hulin, T. Petit, P. Rouanet, J. Jassem, E. Blot, V. Becette, P. Farmer, S. André, C. R. Acharya, S. Mukherjee, D. Cameron, J. Bergh, J. R. Nevins, and R. Iggo. Validation of gene signatures that predict the response of breast cancer to neoadjuvant chemotherapy: a substudy of the eortc 10994/big 00-01 clinical trial. *Lancet Oncology*, 8(12):1071–1078, 2007.

G. Bontempi. *Local Learning Techniques for Modeling, Prediction and Control*. PhD thesis, IRIDIA- Université Libre de Bruxelles, 1999.

A.-L. Boulesteix, C. Porzelius, and M. Daumer. Microarray-based classification and clinical predictors: on combined classifiers and additional predictive value. *Bioinformatics*, 24(15): 1698–1706, 2008. doi: 10.1093/bioinformatics/btn262. URL `http://bioinformatics.oxfordjournals.org/cgi/content/abstract/24/15/169%8`.

L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Chapman and Hall, New-York, 1984.

M. P. S. Brown, W. N. Grundy, D. Lin, N. Cristianini, C. W. Sugnet, T. S. Furey, M. Ares, and D. Haussler. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proceedings of the National Academy of Sciences of the United States of America*, 97(1):262–267, 2000. URL `http://www.pnas.org/content/97/1/262.abstract`.

M. Buyse, S. Loi, L. van't Veer, G. Viale, M. Delorenzi, A. M. Glas, M. Saghatchian d'Assignies, J. Bergh, R. Lidereau, P. Ellis, A. Harris, J. Bogaerts, P. Therasse, A. Floore, M. Amakrane, F. Piette, E. Rutgers, C. Sotiriou, F. Cardoso, and M. J. Piccart. Validation and Clinical Utility of a 70-Gene Prognostic Signature for Women With Node-Negative Breast Cancer. *J. Natl. Cancer Inst.*, 98(17):1183–1192, 2006. doi: 10.1093/jnci/djj329. URL `http://jnci.oxfordjournals.org/cgi/content/abstract/jnci;98/17/1183`.

A. Calabrò, T. Beissbarth, R. Kuner, M. Stojanov, A. Benner, M. Asslaber, F. Ploner, K. Zatloukal, H. Samonigg, A. Poustka, and H. Sültmann. Effects of infiltrating lymphocytes and estrogen receptor on gene expression and prognosis in breast cancer. *Breast Cancer Research and Treatment*, 2008. doi: 10.1007/s10549-008-0105-3. URL `http://dx.doi.org/10.1007/s10549-008-0105-3`.

G. Celeux and G. Govaert. Gaussian parsimonious clustering models. *Pattern Recognition*, 28(5):781–793, 1995.

H. Y. Chang, J. B. Sneddon, A. A. Alizadeh, R. Sood, R. B. West, K. Montgomery, J. Chi, M. van de Rijn, D. Botstein, and P. O. Brown. Gene expression signature of fibroblast serum response predicts human cancer progression: Similarities between tumors and wounds. *PLOS Biology*, 2(2):206–214, 2004.

J. C. Chang, E. C. Wooten, A. Tsimelzon, S. G. Hilsenbeck, M. C. Gutierrez, Y.-L. Tham, M. Kalidas, R. Elledge, S. Mohsin, C. K. Osborne, G. C. Chamness, D. C. Allred, M. T. Lewis, H. Wong, and P. O'Connell. Patterns of Resistance and Incomplete Response to Docetaxel by Gene Expression Profiling in Breast Cancer Patients. *J Clin Oncol*, 23(6): 1169–1177, 2005. doi: 10.1200/JCO.2005.03.156. URL http://jco.ascopubs.org/cgi/content/abstract/23/6/1169.

M. Chanrion, V. Negre, H. Fontaine, N. Salvetat, F. Bibeau, G. M. Grogan, L. Mauriac, D. Katsaros, F. Molina, C. Theillet, and J.-M. Darbon. A Gene Expression Signature that Can Predict the Recurrence of Tamoxifen-Treated Primary Breast Cancer. *Clin Cancer Res*, 14(6):1744–1752, 2008. doi: 10.1158/1078-0432.CCR-07-1833. URL http://clincancerres.aacrjournals.org/cgi/content/abstract/14/6/1744.

S. Chen, S. A. Billings, and W. Luo. Orthogonal least squares methods and their application to non-linear system identification. *International Journal of Control*, 50(5):1873–1896, 1989. URL http://www.informaworld.com/10.1080/00207178908953472.

Y. Cheng and G. M. Church. Biclustering of expression data. *Proc Int Conf Intell Syst Mol Biol*, 8:93–103, 2000. ISSN 1553-0833 (Print).

K. Chin, S. Devries, J. Fridlyand, P. Spellman, R. Roydasgupta, W. L. Kuo, A. Lapuk, R. Neve, Z. Qian, T. Ryder, F. Chen, H. Feiler, T. Tokuyasu, C. Kingsley, S. Dairkee, Z. Meng, K. Chew, D. Pinkel, A. Jain, B. Ljung, L. Esserman, D. Albertson, F. Waldman, and J. Gray. Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. *Cancer cell*, 10:529–41, Dec 2006. URL http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve\&db=pubmed\&%dopt=Abstract\&list_uids=17157792.

E. Chudin, R. Walker, A. Kosaka, S. Wu, D. Rabert, T. Chang, and D. Kreder. Assessment of the relationship between signal intensities and transcript concentration for affymetrix genechip(r) arrays. *Genome Biology*, 3(1):research0005.1–research0005.10, 2001. ISSN 1465-6906. doi: 10.1186/gb-2001-3-1-research0005. URL http://genomebiology.com/2001/3/1/research/0005.

W. G. Cochrane. The combination of estimates from different experiments. *Biometrics*, 10: 101–129, 1954.

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

M. Colozza, E. Azambuja, F. Cardoso, C. Sotiriou, D. Larsimont, and M. J. Piccart. Proliferative markers as prognostic and predictive tools in early breast cancer: where are we now? *Ann Oncol*, 16(11):1723–1739, 2005. doi: 10.1093/annonc/mdi352. URL http://annonc.oxfordjournals.org/cgi/content/abstract/16/11/1723.

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

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

H. Cramer. *Mathematical Methods of Statistics*. Princeton University Press, 1999.

CRAN. Comprehensive R archive network. URL http://cran.r-project.org/.

F. Crick. Central dogma of molecular biology. *Nature*, 227:561–563, August 1970.

N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge University Press, Cambridge, 2000. ISBN 0-521-78019-5.

K. Dahlman-Wright, V. Cavailles, S. A. Fuqua, V. C. Jordan, J. A. Katzenellenbogen, K. S. Korach, A. Maggi, M. Muramatsu, M. G. Parker, and J.-A. Gustafsson. International union of pharmacology. lxiv. estrogen receptors. *Pharmacol Rev*, 58(4):773–781, 2006 Dec. ISSN 0031-6997 (Print). doi: 10.1124/pr.58.4.8.

S. Davies and S. Russell. Np-completeness of searches for smallest possible feature sets. In *AAAI Symposium on Intelligent Relevance*, pages 37–39. AAAI Press, 1994.

C. A. Davis, F. Gerick, V. Hintermair, C. C. Friedel, K. Fundel, R. Kuffner, and R. Zimmer. Reliable gene signatures for microarray classification: assessment of stability and performance. *Bioinformatics*, 22(19):2356–2363, 2006. doi: 10.1093/bioinformatics/btl400. URL http://bioinformatics.oxfordjournals.org/cgi/content/abstract/22/19/235%6.

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

M. de Souto, I. Costa, D. de Araujo, T. Ludermir, and A. Schliep. Clustering cancer gene expression data: a comparative study. *BMC Bioinformatics*, 9(1):497, 2008. ISSN 1471-2105. doi: 10.1186/1471-2105-9-497. URL http://www.biomedcentral.com/1471-2105/9/497.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, B*, 39(1):1–38, 1977.

C. Desmedt, F. Piette, S. Loi, Y. Wang, F. Lallemand, B. Haibe-Kains, G. Viale, M. Delorenzi, Y. Zhang, M. S. d'Assignies, J. Bergh, R. Lidereau, P. Ellis, A. L. Harris, J. G. Klijn, J. A. Foekens, F. Cardoso, M. J. Piccart, M. Buyse, and C. Sotiriou. Strong Time Dependence of the 76-Gene Prognostic Signature for Node-Negative Breast Cancer Patients in the TRANSBIG Multicenter Independent Validation Series. *Clin Cancer Res*, 13(11):3207–3214, 2007. doi: 10.1158/1078-0432.CCR-06-2765. URL http://clincancerres.aacrjournals.org/cgi/content/abstract/13/11/3207.

C. Desmedt, B. Haibe-Kains, P. Wirapati, M. Buyse, D. Larsimont, G. Bontempi, M. Delorenzi, M. Piccart, and C. Sotiriou. Biological Processes Associated with Breast Cancer Clinical Outcome Depend on the Molecular Subtypes. *Clin Cancer Res*, 14(16):5158–5165, 2008. doi: 10.1158/1078-0432.CCR-07-4756. URL http://clincancerres.aacrjournals.org/cgi/content/abstract/14/16/5158.

L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition (Stochastic Modelling and Applied Probability)*. Springer, February 1997. ISBN 0387946187. URL http://www.amazon.ca/exec/obidos/redirect?tag=citeulike09-20\&path=%ASIN/0387946187.

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

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

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

K. Dunne, P. Cunningham, and F. Azuaje. "solutions to instability problems with sequential wrapper-based approaches to feature selection", submitted to the. Technical report, Journal of Machine Learning Research, 2002.

A. Dupuy and R. M. Simon. Critical Review of Published Microarray Studies for Cancer Outcome and Guidelines on Statistical Analysis and Reporting. *J. Natl. Cancer Inst.*, 99 (2):147–157, 2007. doi: 10.1093/jnci/djk018. URL `http://jnci.oxfordjournals.org/cgi/content/abstract/jnci;99/2/147`.

V. Durbecq, J. Toussaint, B. Haibe-Kains, C. Desmedt, G. Rouas, D. Larsimont, M. Buyse, G. Bontempi, M. J. Piccart, and C. Sotiriou. Transforming genomic grade index (GGI) into a user-friendly qrt-pcr tool which will assist clinicians and patients in optimizing treatment of early breast cancer. In J. of Clinical Oncology, editor, *ASCO Annual Meeting Proceedings*, volume 25, page 21058, 2007.

EBCTG. Effects of chemotherapy and hormonal therapy for early breast cancer on recurrence and 15-year survival: an overview of the randomised trials. *Lancet*, 365(9472): 1687–1717, May 2005. ISSN 1474-547X (Electronic). doi: 10.1016/S0140-6736(05)66544-0.

P. Eden, C. Ritz, C. Rose, M. Ferno, and C. Peterson. "good old" clinical markers have similar power in breast cancer prognosis as microarray gene expression profilers. *Eur J Cancer*, 40(12):1837–1841, August 2004. ISSN 0959-8049 (Print). doi: 10.1016/j.ejca.2004.02.025.

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

B. Efron. Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods. *Biometrika*, 68(3):589–599, 1981. doi: 10.1093/biomet/68.3.589. URL `http://biomet.oxfordjournals.org/cgi/content/abstract/68/3/589`.

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

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

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

C. W. Elston and I. O. Ellis. The value of histological grade in breast cancer: experience from a large study with long-term follow-up. *Histopathology*, 19:403–410, 1991.

L. J. Engle, C. L. Simpson, and J. E. Landers. Using high-throughput snp technologies to study cancer. *Oncogene*, 25(11):1594–1601, 2006. URL http://dx.doi.org/10.1038/sj.onc.1209368.

B. S. Everitt. *The Cambridge Dictionary of Statistics*. Cambridge University Press, 2002. ISBN 052181099X.

B. S. Everitt and D. J. Hand. *Finite Mixture Distributions*. Chapman and Hall (London), 1981. ISBN 0-412-22420-8.

G. Finak, N. Bertos, F. Pepin, S. Sadekova, M. Souleimanova, H. Zhao, H. Chen, G. Omeroglu, S. Meterissian, A. Omeroglu, M. Hallett, and M. Park. Stromal gene expression predicts clinical outcome in breast cancer. *Nat Med*, 14(5):518–527, 2008. URL http://dx.doi.org/10.1038/nm1764.

B. Fisher, J. Bryant, N. Wolmark, E. Mamounas, A. Brown, E. Fisher, D. Wickerham, M. Begovic, A. DeCillis, A. Robidoux, R. Margolese, J. Cruz, AB, J. Hoehn, A. Lees, N. Dimitrov, and H. Bear. Effect of preoperative chemotherapy on the outcome of women with operable breast cancer. *J Clin Oncol*, 16(8):2672–2685, 1998. URL http://jco.ascopubs.org/cgi/content/abstract/16/8/2672.

R. A. Fisher. The logic of inductive inference. *Journal of the Royal Statistical Society*, 98: 39–54, 1935.

J. A. Foekens, D. Atkins, Y. Zhang, F. C. Sweep, N. Harbeck, A. Paradiso, T. Cufer, A. M. Sieuwerts, D. Talantov, P. N. Span, V. C. Tjan-Heijnen, A. F. Zito, K. Specht, H. Hioefler, R. Golouh, F. Schittulli, M. Schmitt, L. V. Beex, J. G. Klijn, and Y. Wang. Multicenter validation of a gene expression–based prognostic signature in lymph node–negative primary breast cancer. *Journal of Clinical Oncology*, 24(11), 2006.

C. Fraley and A. E. Raftery. How many clusters? which clustering method? – answers via model-based cluster analysis. *Computational Journal*, 41:578–588, 1998.

J. H. Friedman. On bias, variance, 0/1 loss and the curse of dimensionality. *Data Mining and knowledge Discovery*, pages 564–569, 1996.

M. Friedman. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 32(200):675–701, 1937.

M. H. Galea, R. W. Blamey, C. E. Elston, and I. O. Ellis. The nottingham prognostic index in primary breast cancer. *Breast Cancer Research and Treatment*, 22(3):207–219, 1992.

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

R. Gentleman. Reproducible research: A bioinformatics case study. *Statistical Applications in Genetics and Molecular Biology*, 4(1), 2005.

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

bioinformatics. *Genome Biology*, 5(10):R80, 2004. ISSN 1465-6906. doi: 10.1186/gb-2004-5-10-r80. URL http://genomebiology.com/2004/5/10/R80.

R. Gentleman, W. Huber, V. J. Carey, R. A. Irizarry, and S. Dudoit. *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. Springer, 2005.

T. A. Gerds and M. Schumacher. On functional misspecification of covariates in teh cox regression model. *Biometrika*, 88(2):572–580, 2001.

T. A. Gerds and M. Schumacher. Consistent estimation of the expected brier score in general survival models with right-censored event times. *Biometrical Journal*, 6:1029–1040, 2006.

O. Gevaert, F. D. Smet, D. Timmerman, Y. Moreau, and B. D. Moor. Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks. *Bioinformatics*, 22(14):e184–190, 2006. doi: 10.1093/bioinformatics/btl230. URL http://bioinformatics.oxfordjournals.org/cgi/content/abstract/22/14/e18%4.

M. P. Goetz, V. J. Suman, J. N. Ingle, A. M. Nibbe, D. W. Visscher, C. A. Reynolds, W. L. Lingle, M. Erlander, X.-J. Ma, D. C. Sgroi, E. A. Perez, and F. J. Couch. A Two-Gene Expression Ratio of Homeobox 13 and Interleukin-17B Receptor for Prediction of Recurrence and Survival in Women Receiving Adjuvant Tamoxifen. *Clin Cancer Res*, 12(7):2080–2087, 2006. doi: 10.1158/1078-0432.CCR-05-1263. URL http://clincancerres.aacrjournals.org/cgi/content/abstract/12/7/2080.

A. Goldhirsh, W. C. Wood, R. D. Gelber, A. S. Coates, B. Thurlimann, and H. J. Senn. Meeting highlights: Updated international expert consensus on the primary therapy of early breast cancer. *Journal of Clinical Oncology*, 21(17):3357–3365, 2003.

T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science*, 286(5439):531–537, 1999. doi: 10.1126/science.286.5439.531. URL http://www.sciencemag.org/cgi/content/abstract/286/5439/531.

P. Good. *Resampling Methods*. Birkhauser, third edition, 2006. ISBN 1-8176-4386-9.

E. Graf, C. Schmoor, W. Sauerbrei, and M. Schumacher. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine*, 18:2529–2545, 1999.

B. F. Green. Parameter sensitivity in multivariate methods. *Multivariate Behavioral Research*, 12(3):263–287, 1977. URL http://www.informaworld.com/10.1207/s15327906mbr1203_1.

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

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

J. Gui and H. Li. Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. *Bioinformatics*, 21(13): 3001–3008, 2005. doi: 10.1093/bioinformatics/bti422. URL http://bioinformatics.oxfordjournals.org/cgi/content/abstract/21/13/300%1.

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

B. Haibe-Kains, C. Desmedt, S. Loi, M. Delorenzi, C. Sotiriou, and G. Bontempi. *Computational Intelligence in Clinical Oncology : Lessons Learned from an Analysis of a Clinical Study*, volume 122 of *Studies in Computational Intelligence*, chapter 10, pages 237–268. Springer-Verlag Berlin/Heidelberg, 2008a. ISBN 978-3-540-78533-0. doi: 10.1007/978-3-540-78534-7. URL http://www.springer.com/engineering/book/978-3-540-78533-0.

B. Haibe-Kains, C. Desmedt, F. Piette, M. Buyse, F. Cardoso, L. van't Veer, M. Piccart, G. Bontempi, and C. Sotiriou. Comparison of prognostic gene expression signatures for breast cancer. *BMC Genomics*, 9(1):394, 2008b. ISSN 1471-2164. doi: 10.1186/1471-2164-9-394. URL http://www.biomedcentral.com/1471-2164/9/394.

B. Haibe-Kains, C. Desmedt, C. Sotiriou, and G. Bontempi. A comparative study of survival models for breast cancer prognostication based on microarray data: does a single gene beat them all? *Bioinformatics*, 24(19):2200–2208, 2008c. doi: 10.1093/bioinformatics/btn374. URL http://bioinformatics.oxfordjournals.org/cgi/content/abstract/24/19/220%0.

B. Haibe-Kains, C. Desmedt, F. Rothé, G. Bontempi, and C. Sotiriou. Refining breast cancer prognostication according to the molecular subtypes. *manuscript in preparation*, 2009.

F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel. *Robust Statistics: The Approach Based on Influence Functions (Wiley Series in Probability and Statistics)*. Wiley-Interscience, New York, revised edition, April 2005. ISBN 0471735779. URL http://www.amazon.ca/exec/obidos/redirect?tag=citeulike09-20\&path=%ASIN/0471735779.

D. Hanahan and R. A. Weinberg. The hallmarks of cancer. *Cell*, 100:57–70, 2000.

B. Harr and C. Schlotterer. Comparison of algorithms for the analysis of affymetrix microarray data as evaluated by co-expression of genes in known operons. *Nucleic Acids Research*, 34(2):8, 2006.

F. J. Harrell, K. Lee, and D. Mark. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med*, 15(4):361–387, 1996. doi: 10.1002/(SICI)1097-0258(19960229)15:4⟨361::AID-SIM168⟩3.0.CO;2-4.

J. A. Hartigan and M. A. Wong. A K-means clustering algorithm. *Applied Statistics*, 28: 100–108, 1979.

T. Hastie and R. Tibshirani. *Generalized Additive Models*. Chapman and Hall, London, UK, 1990.

T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning*. Statistics. Springer, 2001. ISBN 978-0-387-95284-0.

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

L. Hedges and I. Olkin. Statistical methods for meta-analysis. *Journal of the American Statistical Association*, 82(397):350–351, 1987. URL http://www.jstor.org/pss/2289186.

K. Hoadley, V. Weigman, C. Fan, L. Sawyer, X. He, M. Troester, C. Sartor, T. Rieger-House, P. Bernard, L. Carey, and C. Perou. Egfr associated expression profiles vary with breast tumor subtype. *BMC Genomics*, 8(1):258, 2007. ISSN 1471-2164. doi: 10.1186/1471-2164-8-258. URL http://www.biomedcentral.com/1471-2164/8/258.

Z. Hu, C. Fan, D. Oh, J. Marron, X. He, B. Qaqish, C. Livasy, L. Carey, E. Reynolds, L. Dressler, A. Nobel, J. Parker, M. Ewend, L. Sawyer, J. Wu, Y. Liu, R. Nanda, M. Tretiakova, A. Orrico, D. Dreher, J. Palazzo, L. Perreard, E. Nelson, M. Mone, H. Hansen, M. Mullins, J. Quackenbush, M. Ellis, O. Olopade, P. Bernard, and C. Perou. The molecular portraits of breast tumors are conserved across microarray platforms. *BMC Genomics*, 7(1):96, 2006. ISSN 1471-2164. doi: 10.1186/1471-2164-7-96. URL http://www.biomedcentral.com/1471-2164/7/96.

D. W. Huang, B. T. Sherman, and R. A. Lempicki. Systematic and integrative analysis of large gene lists using david bioinformatics resources. *Nat. Protocols*, 4(1):44–57, 12 2008. URL http://dx.doi.org/10.1038/nprot.2008.211.

J. Hubble, J. Demeter, H. Jin, M. Mao, M. Nitzberg, T. B. K. Reddy, F. Wymore, Z. K. Zachariah, G. Sherlock, and C. A. Ball. Implementation of GenePattern within the Stanford Microarray Database. *Nucl. Acids Res.*, 37(suppl 1):D898–901, 2009. doi: 10.1093/nar/gkn786. URL http://nar.oxfordjournals.org/cgi/content/abstract/37/suppl_1/D898.

W. Huber, A. von Heydebreck, H. Sultman, A. Poustka, and M. Vingron. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, 18(1):S96–S104, 2002.

C. Ingenuity Systems, Mountain View. Ingenuity pathway analysis. URL http://www.ingenuity.com.

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

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

C. Isaacs, V. Stearns, and D. F. Hayes. New prognostic factors for breast cancer recurrence. *Semin Oncol*, 28(1):53–67, February 2001. ISSN 0093-7754 (Print).

R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton. Adaptive mixtures of local experts. *NC*, 3:79–87, 1991.

A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: A review. *ACM Computing Surveys*, 31(3):264–323, September 1999.

A. Jakulin and I. Bratko. Quantifying and visualizing attribute interactions: An approach based on entropy. *http://arxiv.org/abs/cs.AI/0308002 v3*, 308002:3, 2004.

T. A. Johansen and B. A. Foss. Constructing NARMAX models using ARMAX models. *International Journal of Control*, 58:1125–1153, 1993.

I. T. Jolliffe. *Principal Component Analysis*. Springer Series in Statistics. Springer, 2002. ISBN 978-0-387-95442-4. URL `http://www.springer.com/statistics/statistical+theory+and+methods/book/%978-0-387-95442-4`.

M. J. Jordan and R. A. Jacobs. Hierarchical mixtures of experts and the EM algorithm. *NC*, 6:181–214, 1994.

Kalousis, Alexandros, Prados, Julien, Hilario, and Melanie. Stability of feature selection algorithms: a study on high-dimensional spaces. *Knowledge and Information Systems*, 12(1): 95–116, May 2007. ISSN 0219-1377. doi: http://dx.doi.org/10.1007/s10115-006-0040-8. URL `http://dx.doi.org/10.1007/s10115-006-0040-8`.

A. Kalousis, J. Prados, and M. Hilario. Stability of feature selection algorithms. In *ICDM '05: Proceedings of the Fifth IEEE International Conference on Data Mining*, pages 218–225, Washington, DC, USA, 2005. IEEE Computer Society. ISBN 0-7695-2278-5. doi: http://dx.doi.org/10.1109/ICDM.2005.135.

M. D. Kane, T. A. Jatkoe, C. R. Stumpf, J. Lu, J. D. Thomas, and S. J. Madore. Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays. *Nucleic Acids Res*, 28(22):4552–4557, November 2000. ISSN 1362-4962 (Electronic).

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

A. Kapp, S. Jeffrey, A. Langerod, A.-L. Borresen-Dale, W. Han, D.-Y. Noh, I. Bukholm, M. Nicolau, P. Brown, and R. Tibshirani. Discovery and validation of breast cancer subtypes. *BMC Genomics*, 7(1):231, 2006. ISSN 1471-2164. doi: 10.1186/1471-2164-7-231. URL `http://www.biomedcentral.com/1471-2164/7/231`.

M. W. Kattan. Evaluating a New Marker's Predictive Contribution. *Clin Cancer Res*, 10 (3):822–824, 2004. doi: 10.1158/1078-0432.CCR-03-0061. URL `http://clincancerres.aacrjournals.org`.

J. Kittler, M. Hatef, R. Duin, and J. Matas. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–238, 1998.

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

G. E. Konecny, Y. G. Meng, M. Untch, H.-J. Wang, I. Bauerfeind, M. Epstein, P. Stieber, J.-M. Vernes, J. Gutierrez, K. Hong, M. Beryt, H. Hepp, D. J. Slamon, and M. D. Pegram. Association between her-2/neu and vascular endothelial growth factor expression predicts

clinical outcome in primary breast cancer patients. *Clin Cancer Res*, 10(5):1706–1716, March 2004. ISSN 1078-0432 (Print).

E. L. Korn, J. Troendie, L. M. McShane, and R. Simon. Controlling the number of false discoveries: Application to high dimensional genomic data. *Journal of Statist Plann Inference*, 124:379–398, 2004.

P. Krizek. *Feature selection: stability, algorithms, and evaluation*. PhD thesis, Czech Technical University in Prague, June 2008.

L. Leon, T. Cai, and L. J. Wei. Robust inferences for covariate effects on survival time with censored linear regression models. Technical Report Working Paper 20, Harvard University, January 2005.

S. Lewis and M. Clarke. Forest plots: trying to see the wood and the trees. *British Medical Journal*, 322(7300):1479–1480, 2001.

C. Li and W. H. Wong. Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. *Genome Biology*, 2(8):1–11, 2001.

D. Y. Lin and J. Weil. The robust inference for the cox proportional hazards model. *Journal of American Statistical Asscoiation*, 84(408):1074–1078, 1989.

L.-T. Liu, J.-P. Peng, H.-C. Chang, and W.-C. Hung. Reck is a target of epstein-barr virus latent membrane protein 1. *Oncogene*, 22(51):8263–8270, 2003. URL `http://dx.doi.org/10.1038/sj.onc.1207157`.

R. Liu, X. Wang, G. Y. Chen, P. Dalerba, A. Gurney, T. Hoey, G. Sherlock, J. Lewicki, K. Shedden, and M. F. Clarke. The prognostic role of a gene signature from tumorigenic breast-cancer cells. *The New England Journal of Medicine*, 356(3):217–226, 2007.

S. Loi, B. Haibe-Kains, C. Desmedt, F. Lallemand, A. M. Tutt, C. Gillet, P. Ellis, A. Harris, J. Bergh, J. A. Foekens, J. G. Klijn, D. Larsimont, M. Buyse, G. Bontempi, M. Delorenzi, M. J. Piccart, and C. Sotiriou. Definition of Clinically Distinct Molecular Subtypes in Estrogen Receptor-Positive Breast Carcinomas Through Genomic Grade. *J Clin Oncol*, 25 (10):1239–1246, 2007. doi: 10.1200/JCO.2006.07.1522. URL `http://jco.ascopubs.org/cgi/content/abstract/25/10/1239`.

S. Loi, B. Haibe-Kains, C. Desmedt, P. Wirapati, F. Lallemand, A. Tutt, C. Gillet, P. Ellis, K. Ryder, J. Reid, M. Daidone, M. Pierotti, E. Berns, M. Jansen, J. Foekens, M. Delorenzi, G. Bontempi, M. Piccart, and C. Sotiriou. Predicting prognosis using molecular profiling in estrogen receptor-positive breast cancer treated with tamoxifen. *BMC Genomics*, 9(1):239, 2008. ISSN 1471-2164. doi: 10.1186/1471-2164-9-239. URL `http://www.biomedcentral.com/1471-2164/9/239`.

P. Lonning, S. Knappskog, V. Staalesen, R. Chrisanthar, and J. Lillehaug. Breast cancer prognostication and prediction in the postgenomic era. *Ann Oncol*, 18(8):1293–1306, 2007. doi: 10.1093/annonc/mdm013. URL `http://annonc.oxfordjournals.org/cgi/content/abstract/18/8/1293`.

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

J. B. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297, 1967.

D. Mauri, N. Pavlidis, and J. P. A. Ioannidis. Neoadjuvant Versus Adjuvant Systemic Treatment in Breast Cancer: A Meta-Analysis. *J. Natl. Cancer Inst.*, 97(3):188–194, 2005. doi: $10.1093/\mathrm{jnci/dji021}$. URL `http://jnci.oxfordjournals.org/cgi/content/abstract/jnci;97/3/188`.

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

L. M. McShane, D. G. Altman, W. Sauerbrei, S. E. Taube, M. Gion, and G. M. Clark. Reporting Recommendations for Tumor Marker Prognostic Studies (REMARK). *J. Natl. Cancer Inst.*, 97(16):1180–1184, 2005. doi: $10.1093/\mathrm{jnci/dji237}$. URL `http://jnci.oxfordjournals.org/cgi/content/abstract/jnci;97/16/1180`.

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

P. E. Meyer. *Information-Theoretic Variable Selection and Network Inference from Microarray Data*. PhD thesis, Université Libre de Bruxelles, December 2008.

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

L. D. Miller, P. M. Long, L. Wong, S. Mukherjee, L. M. McShane, and E. T. Liu. Optimal gene expression analysis by microarrays. 2(5):353–361, 11 2002. URL `http://linkinghub.elsevier.com/retrieve/pii/S1535610802001812`.

L. D. Miller, J. Smeds, J. George, V. B. Vega, L. Vergara, A. Pioner, Y. Pawitan, P. Hall, S. Klaar, E. T. Liu, and J. Bergh. An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *PNAS*, 102 (38):13550–13555, 2005.

A. J. Minn, G. P. Gupta, P. M. Siegel, P. D. Bos, W. Shu, D. D. Giri, A. Viale, A. B. Olshen, W. L. Gerald, and J. Massague. Genes that mediate breast cancer metastasis to lung. *Nature*, 436(7050):518–524, 2005. URL `http://dx.doi.org/10.1038/nature03799`.

A. J. Minn, G. P. Gupta, D. Padua, P. Bos, D. X. Nguyen, D. Nuyten, B. Kreike, Y. Zhang, Y. Wang, H. Ishwaran, J. A. Foekens, M. van de Vijver, and J. Massague. Lung metastasis genes couple breast tumor size and metastatic spread. *Proceedings of the National Academy of Sciences*, 104(16):6740–6745, 2007. doi: $10.1073/\mathrm{pnas.0701138104}$. URL `http://www.pnas.org/cgi/content/abstract/104/16/6740`.

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

A. M. Molinaro, S. Dudoit, and v. M. J. Tree-based multivariate regresion and density estimation with right-censored data. *Journal of Multivariate Analaysis*, 90:154–177, 2004.

J. Moody and C. J. Darken. Fast learning in networks of locally-tuned processing units. *Neural Computation*, 1(2):281–294, 1989.

D. Murphy. Gene Expression Studies Using Microarrays: Principles, Problems and Prospects. *Advan. Physiol. Edu.*, 26(4):256–270, 2002. URL `http://advan.physiology.org/cgi/content/abstract/26/4/256`.

R. Murray-Smith. *A local model network approach to nonlinear modelling*. PhD thesis, Department of Computer Science, University of Strathclyde, Strathclyde, UK, 1994.

R. Murray-Smith and T. A. Johansen. Local learning in local model networks. In R. Murray-Smith and T. A. Johansen, editors, *Multiple Model Approaches to Modeling and Control*, chapter 7, pages 185–210. Taylor and Francis, 1997.

R. H. Myers. *Classical and Modern Regression with Applications*. PWS-KENT Publishing Company, Boston, MA, second edition, 1994.

A. Naderi, A. E. Teschendorff, N. L. Barbosa-Morais, S. E. Pinder, A. R. Green, D. G. P. andJ . F. Robertson, S. Aparicio, I. O. Ellis, J. D. Brenton, and C. Caldas. A gene-expression signature to predict survival in breast cancer across independent data sets. *Oncogene*, 26:1507–1516, 2007.

E. Nimeus-Malmstrom, M. Krogh, P. Malmstrom, C. Strand, I. Fredriksson, P. Karlsson, B. Nordenskjold, O. Stal, G. Ostberg, C. Peterson, and M. Ferno. Gene expression profiling in primary breast cancer distinguishes patients developing local recurrence after breast-conservation surgery, with or without postoperative radiotherapy. *Breast Cancer Research*, 10(2):R34, 2008. ISSN 1465-5411. doi: $10.1186/\mathrm{bcr1997}$. URL `http://breast-cancer-research.com/content/10/2/R34`.

I. A. Olivotto, C. D. Bajdik, P. M. Ravdin, C. H. Speers, A. J. Coldman, B. D. Norris, G. J. Davis, S. K. Chia, and K. A. Gelmon. Population-based validation of the prognostic model adjuvant! for early breast cancer. *Journal of Clinical Oncology*, 23(12):2716–2725, 2005.

A. B. Owen, J. Stuart, K. Mach, A. M. Villeneuve, and S. Kim. A gene recommender algorithm to identify coexpressed genes in c. elegans. *Genome Res*, 13(8):1828–1837, August 2003. ISSN 1088-9051 (Print). doi: $10.1101/\mathrm{gr.1125403}$.

S. Paik, S. Shak, G. Tang, C. Kim, J. Bakker, M. Cronin, F. L. Baehner, M. G. Walker, D. Watson, T. Park, W. Hiller, E. R. Fisher, D. L. Wickerham, J. Bryant, and N. Wolmark. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *New England Journal of Medicine*, 351(27):2817–2826, 2004.

M. Park, T. Hastie, and R. Tibshirani. Averaged gene expression for regression. *Biostatistics*, 8(2):212–227, 2007.

D. M. Parkin, F. Bray, J. Ferlay, and P. Pisani. Estimating the world cancer burden: Globocan 2000. *Int J Cancer*, 94(2):153–156, October 2001. ISSN 0020-7136 (Print).

H. Parkinson, U. Sarkans, a. A. M. Shojatalab, S. Contrino, R. Coulson, A. Farne, G. G. Lara, E. Holloway, M. Kapushesky, P. Lilja, G. Mukherjee, A. Oezcimen, T. Rayner, P. Rocca-Serra, A. Sharma, S. Sansone, and A. Brazma. Arrayexpress: a public repository for microarray gene expression data at the ebi. *Nucleic Acids Research*, 33:D553–D555, 2005.

Y. Pawitan, J. Bjohle, L. Amler, A. Borg, S. Egyhazi, P. Hall, X. Han, L. Holmberg, F. Huang, S. Klaar, E. T. Liu, L. Miller, H. Nordgren, A. Ploner, K. Sandelin, P. M. Shaw, J. Smeds, L. Skoog, S. Wedren, and J. Bergh. Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts. *Breast Cancer Research*, 7(6):953–964, 2005.

A. C. Pease, D. Solas, E. J. Sullivan, M. T. Cronin, C. P. Holmes, and S. P. Fodor. Light-generated oligonucleotide arrays for rapid dna sequence analysis. *Proc Natl Acad Sci USA*, 91(11):5022–5026, May 1994. ISSN 0027-8424 (Print).

M. Pencina and R. D'Agostino. Overall c as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. *Stat Med*, 23(13): 2109–2123, 2004. doi: $10.1002/\mathrm{sim}.1802$.

E. A. Perez, V. J. Suman, N. E. Davidson, S. Martino, P. A. Kaufman, W. L. Lingle, P. J. Flynn, J. N. Ingle, D. Visscher, and R. B. Jenkins. HER2 Testing by Local, Central, and Reference Laboratories in Specimens From the North Central Cancer Treatment Group N9831 Intergroup Adjuvant Trial. *J Clin Oncol*, 24(19):3032–3038, 2006. doi: $10.1200/\mathrm{JCO}.2005.03.4744$. URL `http://jco.ascopubs.org/cgi/content/abstract/24/19/3032`.

C. M. Perou, T. Sorlie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, C. A. Rees, J. R. Pollack, D. T. Ross, H. Johnsen, L. A. Akslen, O. Fluge, A. Pergamenschikov, C. Williams, S. X. Zhu, P. E. Lonning, A.-L. Borresen-Dale, P. O. Brown, and D. Botstein. Molecular portraits of human breast tumours. *Nature*, 406(6797):747–752, 2000. URL `http://dx.doi.org/10.1038/35021093`.

M. P. Perrone and L. N. Cooper. When networks disagree: Ensemble methods for hybrid neural networks. In R. J. Mammone, editor, *Artificial Neural Networks for Speech and Vision*, pages 126–142. Chapman and Hall, 1993.

R. L. Plackett. Karl pearson and the chi-squared test. *International Statistical Review*, 51(1): 59–72, 1983.

A. Ploner, L. D. Miller, P. Hall, J. Bergh, and Y. Pawitan. Correlation test to assess low-level processing of high-density oligonucletide microarray data. *BMC Bioinformatics*, 6(80):1–20, 2005.

A. Prasad, A. Z. Fernandis, Y. Rao, and R. K. Ganju. Slit protein-mediated inhibition of cxcr4-induced chemotactic and chemoinvasive signaling pathways in breast cancer cells. *J Biol Chem*, 279(10):9115–9124, March 2004. ISSN 0021-9258 (Print). doi: $10.1074/\mathrm{jbc}.\mathrm{M308083200}$.

V. Praz, V. Jagannathan, and P. Bucher. Cleanex: a database of heterogeneous gene expression data based on a consistent gene nomenclature. *Nucleic Acids Research*, 32: 542–547, 2003.

L. Pusztai, C. Mazouni, K. Anderson, Y. Wu, and W. F. Symmans. Molecular Classification of Breast Cancer: Limitations and Potential. *Oncologist*, 11(8):868–877, 2006. doi: 10.1634/theoncologist.11-8-868. URL http://theoncologist.alphamedpress.org/cgi/content/abstract/11/8/868.

R Development Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2007. URL http://www.R-project.org. ISBN 3-900051-07-0.

R Development Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org.

S. Ramaswamy, K. N. Ross, E. S. Lander, and T. R. Golub. A molecular signature of metastasis in primary solid tumors. *Nature Genetics*, 33:49–53, 2003.

P. M. Ravdin, L. A. Siminoff, G. J. Davis, M. B. Mercer, J. Hewlett, N. Gerson, and H. L. Parker. Computer Program to Assist in Making Decisions About Adjuvant Therapy for Women With Early Breast Cancer. *J Clin Oncol*, 19(4):980–991, 2001. URL http://jco.ascopubs.org/cgi/content/abstract/19/4/980.

J. F. Reid, L. Lusa, L. De Cecco, D. Coradini, S. Veneroni, M. G. Daidone, M. Gariboldi, and M. A. Pierotti. Limits of Predictive Models Using Microarray Data for Breast Cancer Clinical Treatment Outcome. *J. Natl. Cancer Inst.*, 97(12):927–930, 2005. doi: 10.1093/jnci/dji153. URL http://jnci.oxfordjournals.org/cgi/content/abstract/jnci;97/12/927.

D. R. Rhodes, S. Kalyana-Sundaram, V. Mahavisno, R. Varambally, J. Yu, B. B. Briggs, T. R. Barrette, M. J. Anstet, C. Kincead-Beal, P. Kulkarni, S. Varambally, D. Ghosh, and A. M. Chinnaiyan. Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles. *Neoplasia*, 9(2):166–180, February 2007. ISSN 1476-5586 (Electronic).

R. M. Ripley, A. L. Harris, and L. Tarassenko. Non-linear survival analysis using neural networks. *Statistics in Medicine*, 23:825–842, 2004. doi: 10.1002/sim.1655.

J. L. Rodgers and A. W. Nicewander. Thirteen ways to look at the correlation coefficient. *The American Statistician*, 42(1):59–66, 1988. doi: http://dx.doi.org/10.2307/2685263. URL http://dx.doi.org/10.2307/2685263.

P. J. Rousseeuw and A. M. Leroy. *Robust regression and outlier detection*. John Wiley & Sons, Inc., New York, NY, USA, 2003.

P. J. Rousseuw. Least median squares regression. *Journal of American Statistical Asscoiation*, 79:871–880, 1984.

R. Rouzier, C. M. Perou, W. F. Symmans, N. Ibrahim, M. Cristofanilli, K. Anderson, K. R. Hess, J. Stec, M. Ayers, P. Wagner, P. Morandi, C. Fan, I. Rabiul, J. S. Ross, G. N. Hortobagyi, and L. Pusztai. Breast Cancer Molecular Subtypes Respond Differently to Preoperative Chemotherapy. *Clin Cancer Res*, 11(16):5678–5685, 2005. doi: 10.1158/1078-0432.CCR-04-2421. URL http://clincancerres.aacrjournals.org/cgi/content/abstract/11/16/5678.

P. Royston and W. Sauerbrei. A new measure of prognostic separation in survival data. *Statistics in Medicine*, 23:723–748, 2004.

P. Royston, D. G. Altman, and W. Sauerbrei. Dichotomizing continuous predictors in multiple regression: a bad idea. *Stat Med*, 25(1):127–141, January 2006. ISSN 0277-6715 (Print). doi: 10.1002/sim.2331.

L. H. Saal, P. Johansson, K. Holm, S. K. Gruvberger-Saal, Q.-B. She, M. Maurer, S. Koujak, A. A. Ferrando, P. Malmstrom, L. Memeo, J. Isola, P.-O. Bendahl, N. Rosen, H. Hibshoosh, M. Ringner, A. Borg, and R. Parsons. Poor prognosis in carcinoma is associated with a gene expression signature of aberrant PTEN tumor suppressor pathway activity. *Proceedings of the National Academy of Sciences*, 104(18):7564–7569, 2007. doi: 10.1073/pnas.0702507104. URL http://www.pnas.org/content/104/18/7564.abstract.

Y. Saeys, I. Inza, and P. Larranaga. A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19):2507–2517, 2007. doi: 10.1093/bioinformatics/btm344. URL http://bioinformatics.oxfordjournals.org/cgi/content/abstract/23/19/250%7.

R. W. Scarff and H. Torloni. Histological typing of breast tumors. *International histological classification of tumours*, 2(2):13–20, 1968.

M. Schmidt, D. Bohm, C. von Torne, E. Steiner, A. Puhl, H. Pilch, H.-A. Lehr, J. G. Hengstler, H. Kolbl, and M. Gehrmann. The Humoral Immune System Has a Key Prognostic Impact in Node-Negative Breast Cancer. *Cancer Res*, 68(13):5405–5413, 2008. doi: 10.1158/0008-5472.CAN-07-5206. URL http://cancerres.aacrjournals.org/cgi/content/abstract/68/13/5405.

G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.

Q. Sheng, Y. Moreau, and B. De Moor. Biclustering microarray data by Gibbs sampling. *Bioinformatics*, 19(suppl 2):ii196–205, 2003. doi: 10.1093/bioinformatics/btg1078. URL http://bioinformatics.oxfordjournals.org/cgi/content/abstract/19/suppl_%2/ii196.

Q. Sheng, Y. Moreau, F. De Smet, K. Marchal, and B. De Moor. *Data Analysis and Visualization in Genomics and Proteomics*. John Wiley & Sons, 2005. ISBN 9780470094396. URL http://www3.interscience.wiley.com/cgi-bin/summary/110528999/SUMMARY.

R. Simon. Diagnostic and prognostic prediction using gene expression profiles in high-dimensional microarray data. *British Journal of Cancer*, 89:1599–1604, 2003.

R. Simon, M. D. Radmacher, K. Dobbin, and L. M. McShane. Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *J. Natl Cancer Inst*, 95(1):14–18, 2003.

S. E. Singletary, C. Allred, P. Ashley, L. W. Bassett, D. Berry, K. I. Bland, P. I. Borgen, G. Clark, S. B. Edge, D. F. Hayes, L. L. Hughes, R. V. Hutter, M. Morrow, D. L. Page, A. Recht, R. L. Theriault, A. Thor, D. L. Weaver, H. S. Wieand, and F. L. Greene. Revision of the American Joint Committee on Cancer Staging System for Breast Cancer. *J Clin Oncol*, 20(17):3628–3636, 2002. doi: 10.1200/JCO.2002.02.026. URL http://jco.ascopubs.org/cgi/content/abstract/20/17/3628.

T. Sorlie, C. M. Perou, R. Tibshirani, T. Aas, S. Geisher, H. Johnsen, T. Hastie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, T. Thorsen, H. Quist, J. C. Matese., P. O. Brown, D. Botstein, P. L. Eystein, and A. L. Borresen-Dale. Gene expression patterns breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Matl. Acad. Sci. USA*, 98 (19):10869–10874, 2001.

T. Sorlie, R. Tibshirani, J. Parker, T. Hastie, J. S. Marron, A. Nobel, S. Deng, H. Johnsen, R. Pesich, S. Geister, J. Demeter, C. Perou, P. E. Lonning, P. O. Brown, A. L. Borresen-Dale, and D. Botstein. Repeated observation of breast tumor subtypes in indepedent gene expression data sets. *Proc Natl Acad Sci USA*, 1(14):8418–8423, 2003.

C. Sotiriou and M. J. Piccart. Taking gene-expression profiling to the clinic: when will molecular signatures become relevant to patient care? *Nature Cancer Review*, 7:545–553, July 2007.

C. Sotiriou and L. Pusztai. Gene-Expression Signatures in Breast Cancer. *N Engl J Med*, 360(8):790–800, 2009. doi: $10.1056/\mathrm{NEJMra0801289}$. URL `http://content.nejm.org`.

C. Sotiriou, S. Y. Neo, L. M. McShane, E. L. Korn, P. M. Long, A. Jazaeri, P. Martiat, S. Fox, A. L. Harris, and E. T. Liu. Breast cancer classification and prognosis based on gene expression profiles from a population-based study. *Proc. Natl. Acad. Sci.*, 100(18):10393–10398, 2003.

C. Sotiriou, P. Wirapati, S. Loi, A. Harris, S. Fox, J. Smeds, H. Nordgren, P. Farmer, V. Praz, B. Haibe-Kains, C. Desmedt, D. Larsimont, F. Cardoso, H. Peterse, D. Nuyten, M. Buyse, M. J. Van de Vijver, J. Bergh, M. Piccart, and M. Delorenzi. Gene Expression Profiling in Breast Cancer: Understanding the Molecular Basis of Histologic Grade To Improve Prognosis. *J. Natl. Cancer Inst.*, 98(4):262–272, 2006. doi: $10.1093/\mathrm{jnci/djj052}$. URL `http://jnci.oxfordjournals.org/cgi/content/abstract/jnci;98/4/262`.

P. Sprites, C. Glymour, and R. Scheines. *Causation, prediction, and search: Adaptive computation and machine learning*. MIT Press, second edition, 2000.

J. C. Stone. Additive regression and other nonparametric models. *Annals of Statistics*, 13: 689–705, 1985.

M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society B*, 36(1):111–147, 1974.

C. Sugar. *Techniques for clustering and classification with applications to medical problems*. PhD thesis, Stanford University, 1998.

J. A. Sweets. Measuring the accuracy of diagnostic systems. *Science*, 240(4857):1285–1293, 1988.

P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E. S. Lander, and T. R. Golub. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc Natl Acad Sci U S A*, 96(6):2907–2912, March 1999. ISSN 0027-8424 (Print).

A. Teschendorff and C. Caldas. A robust classifier of high predictive value to identify good prognosis patients in er-negative breast cancer. *Breast Cancer Research*, 10(4):R73, 2008. ISSN 1465-5411. doi: 10.1186/bcr2138. URL http://breast-cancer-research.com/content/10/4/R73.

A. Teschendorff, A. Miremadi, S. Pinder, I. Ellis, and C. Caldas. An immune response gene expression module identifies a good prognosis subtype in estrogen receptor negative breast cancer. *Genome Biology*, 8(8):R157, 2007. ISSN 1465-6906. doi: 10.1186/gb-2007-8-8-r157. URL http://genomebiology.com/2007/8/8/R157.

The Tumor Analysis Best Practices Working Group. Expression profiling - best practices for data generation and interpretation in clinical trials. *Nat Rev Genet*, 5(3):229–237, 03 2004. URL http://dx.doi.org/10.1038/nrg1297.

T. M. Therneau and P. M. Grambsch. *Modeling Survival Data: Extending the Cox Model*. Statistics for Biology and Health. Springer-Verlag New-York, 2000. ISBN 978-0-387-98784-2. doi: 10.1002/sim.956. URL http://www.springer.com/statistics/stats+life+sci/book/978-0-387-98784-%2.

R. Tibshirani and G. Walther. Cluster validation by prediction strength. *Journal of Computational and Graphical Statistics*, 14(3):511–528, 2005.

R. Tibshirani, G. Walther, and T. Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423, 2001. doi: 10.1111/1467-9868.00293.

R. J. Tibshirani and B. Efron. Pre-validation and inference in microarrays. *Stat Appl Genet Mol Biol*, 1:Article1, 2002. ISSN 1544-6115 (Electronic). doi: 10.2202/1544-6115.1000.

J. H. Todd, C. Dowle, M. R. Wiliam, C. W. Elston, I. O. Ellis, C. P. Hinton, R. W. Blamey, and J. L. Haybittle. Confirmation of a prognostic index in primary breast cancer. *British Journal of Cancer*, 56(4):489–492, 1987.

J. Toussaint, A. Sieuwerts, V. Durbecq, B. Haibe-Kains, E. Berns, A. L. Harris, D. Larsimont, M. Piccart, J. Foekens, and C. Sotiriou. Molecular qrt-pcr grade index: a new tool for breast cancer (bc) patient grading improvement. *European Journal of Cancer Supplements*, 6(7): 140, April 2008.

G. C. Tseng and W. H. Wong. Tight clustering: a resampling-based approach for identifying stable and tight patterns in data. *Biometrics*, 61(1):10–16, March 2005. ISSN 0006-341X (Print). doi: 10.1111/j.0006-341X.2005.031032.x.

S. Turner, J. A Sherratt, and D. Cameron. Tamoxifen treatment failure in cancer and the nonlinear dynamics of tgfbeta. *J Theor Biol*, 229(1):101–111, July 2004. ISSN 0022-5193 (Print). doi: 10.1016/j.jtbi.2004.03.008.

P. Turney. Technical note: Bias and the quantification of stability. *Mach. Learn.*, 20(1-2): 23–33, 1995. ISSN 0885-6125. doi: http://dx.doi.org/10.1007/BF00993473.

V. van Belle, K. Pelckmans, J. A. Suykens, and S. van Huffel. Support vector machines for survival analysis. In *Third International Conference on Computational Intelligence in Medicine and Healthcare*, 2007.

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

M. J. van der Laan, K. S. Pollard, and J. Bryan. A new partitioning around medoids algorithm. *Journal of Statistical Computation and Simulation*, 73(8):575–584, 2003.

H. van Houwelingen, T. Bruinsma, A. A. Hart, L. J. van't Veer, and L. F. A. Wessels. Cross-validated cox regression on microarray gene expression data. *Statistics in Medicine*, 25: 3201–3216, 2006.

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

U. Veronesi, P. Boyle, A. Goldhirsch, R. Orecchia, and G. Viale. Breast cancer. *Lancet*, 365 (9472):1727–1741, May 2005. ISSN 1474-547X (Electronic). doi: 10.1016/S0140-6736(05) 66546-4.

P. J. M. Verweij and J. C. van Houwelingen. Cross-validation in survival analysis. *Statistics in Medicine*, 12:2305–2314, 1993.

G. A. Viswanathan, J. Seto, S. Patil, G. Nudelman, and S. C. Sealfon. Getting started in biological pathway construction and analysis. *PLoS Comput Biol*, 4(2):e16, Feb 2008. doi: 10.1371/journal.pcbi.0040016. URL `http://dx.doi.org/10.1371%2Fjournal.pcbi.0040016`.

H. Wainer. Estimating coefficients in linear models: It don't make no nevermind. *Psychological Bulletin*, 83(2):213–217, 1976.

D. Wang and S. J. Lippard. Cellular processing of platinum anticancer drugs. *Nat Rev Drug Discov*, 4(4):307–320, 2005. URL `http://dx.doi.org/10.1038/nrd1691`.

Y. Wang, J. G. Klijn, Y. Zhang, A. M. Sieuwerts, M. P. Look, F. Yang, D. Talantov, M. Timmermans, M. E. M. van Gelder, J. Yu, T. Jatkoe, E. M. Berns, D. Atkins, and J. A. Forekens. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet*, 365:671–679, 2005.

L. Wasserman. *All of Nonparametric Statistics*. Springer, 2007. ISBN 0387251456.

A. Webb. *Statistical Pattern Recognition*. John Wiley and Sons, second edition, 2003. ISBN 0-470-84513-9.

E. Werner. Genome semantics, in silico multicellular systems and the central dogma. *FEBS Lett*, 579(8):1779–1782, March 2005. ISSN 0014-5793 (Print). doi: 10.1016/j.febslet.2005.02.011.

L. F. A. Wessels, M. J. T. Reinders, T. van Welsem, and P. M. Nederlof. Representation and classification for high-throughput data. In *Proceedings of SPIE*, volume 4626, San Jose, California, USA, 2002. BIOS2002.

L. F. A. Wessels, M. J. T. Reinders, A. A. M. Hart, C. J. Veenman, H. Dai, Y. D. He, and L. J. v. Veer. A protocol for building and evaluating predictors of disease state based on microarray data. *Bioinformatics*, 21(19):3755–3762, 2005. doi: 10.1093/bioinformatics/bti429. URL http://bioinformatics.oxfordjournals.org/cgi/content/abstract/21/19/375%5.

W. Wienholt and B. Sendhoff. How to determine the redundancy of noisy chaotic time series, 1996. URL citeseer.comp.nus.edu.sg/150922.html.

Wikipedia. The free encyclopedia. URL http://www.wikipedia.org/.

F. Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1:80–83, 1945.

P. Wirapati, C. Sotiriou, S. Kunkel, P. Farmer, S. Pradervand, B. Haibe-Kains, C. Desmedt, M. Ignatiadis, T. Sengstag, F. Schutz, D. Goldstein, M. Piccart, and M. Delorenzi. Meta-analysis of gene expression profiles in breast cancer: toward a unified understanding of breast cancer subtyping and prognosis signatures. *Breast Cancer Research*, 10(4):R65, 2008. ISSN 1465-5411. doi: 10.1186/bcr2124. URL http://breast-cancer-research.com/content/10/4/R65.

S. H. Woolf. The Meaning of Translational Research and Why It Matters. *JAMA*, 299(2): 211–213, 2008. doi: 10.1001/jama.2007.26. URL http://jama.ama-assn.org.

G. Wright, B. Tan, A. Rosenwald, E. H. Hurt, A. Wiestner, and L. M. Staudt. A gene expression-based method to diagnose clinically distinct subgroups of diffuse large b-cell lymphoma. *PNAS*, 100(17):9991–9996, 2003.

Z. Wu and R. A. Irizarry. Preprocessing of oligonucleotide array data. *Nature Biotechnology*, 22:656–658, 2004.

K. Y. Yeung, C. Fraley, A. M. an A. E. Raftery, and W. L. Ruzzo. Model-based clustering and data transformations for gene expression data. *Bioinformatics*, 17(10):977–987, 2001.

B. Zupan, J. Demšar, M. Kattan, J. Beck, and I. Bratko. Machine learning for survival analysis: A case study on recurrence of prostate cancer. *Artificial Intelligence in Medicine*, pages 346–355, 1999. URL http://dx.doi.org/10.1007/3-540-48720-4_37.

Appendix A

PRESS

Cross-validation provides a reliable estimate of the generalization error of a predictive model [Stone, 1974]. The disadvantage of such an approach is that it requires the training process to be repeated multiple times, which may be computational intensive (Figure A.1 (a)). However, in the case of linear models, there exists a powerful statistical procedure to compute the leave-one-out cross-validation (LOOCV) errors (or *residuals*) at a reduced computational cost (Figure A.1 (b)). It is the PRESS (prediction sum of squares) statistic [Allen, 1974], a simple formula which returns the LOOCV residuals as a byproduct of the parametric identification of β in a linear model.

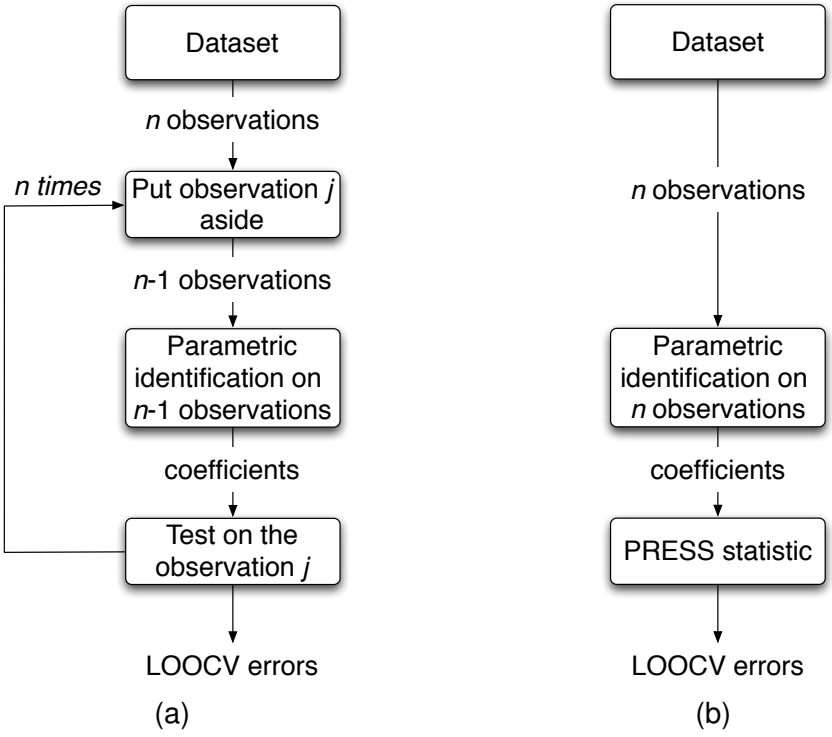


Figure A.1: (a) LOOCV procedure; (b) PRESS statistic.

Let $D = \langle X, y \rangle$ be a training dataset where X is the matrix of $p+1$ independent variables and y is the vector of the dependent variable for n individuals. In this section, we consider

the linear regression model

$$y = x^T \beta$$

where x stands or the vector of $p + 1$ values

$$x = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix}$$

Consider D in which for n times

1. We set aside the j^{th} observation from the training set D.
2. We use the remaining $n - 1$ observations to estimate the linear regression coefficients $\hat{\beta}^{-j}$.
3. We use $\hat{\beta}^{-j}$ to predict the dependent variable $\hat{y}_j^{-j}$ for x_j.

The LOOCV residual for the j^{th} observation is

$$e_j^{loocv} = y_j - \hat{y}_j^{-j} = y_j - x_j^T \hat{\beta}^{-j}$$

The PRESS statistic is an efficient way to compute the LOOCV residuals on the basis of a simple regression performed on the whole training set. This enables a fast cross-validation without repeating n times the training procedure. The PRESS procedure is the following:

1. We use the whole training set to estimate the linear regression coefficients $\hat{\beta}$. This procedure is performed only once on the n observations and returns as byproduct the *Hat* matrix [Myers, 1994]

$$H = X (X^T X)^{-1} X^T \tag{A.1}$$

2. We compute the vector of residuals e whose j^{th} term is $e_j = y_j - x_j^T \hat{\beta}$
3. We use the PRESS statistic to compute the e_j^{loocv} as

$$e_j^{loocv} = \frac{e_j}{1 - H_{jj}}$$

So the vector e^{loocv} is the vector of LOOCV errors of the linear model $y = x^T \beta$, as computed by the PRESS statistic.

Appendix B

Experimental Findings

B.1 Breast Cancer Molecular Subtypes

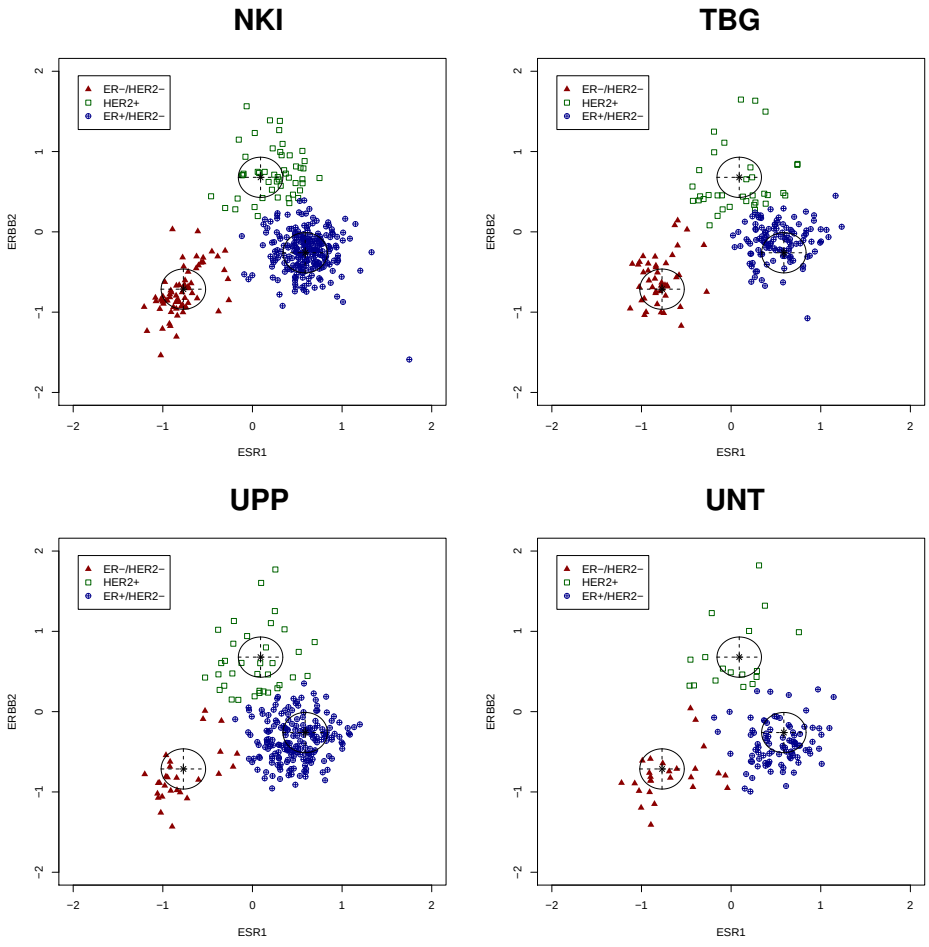


Figure B.1: Classification of the tumors using the subtype clustering model in the NKI, TBG, UPP and UNT datasets. Each subtype is represented by a different color and symbol. The superimposed ellipses correspond to the covariance of the components.

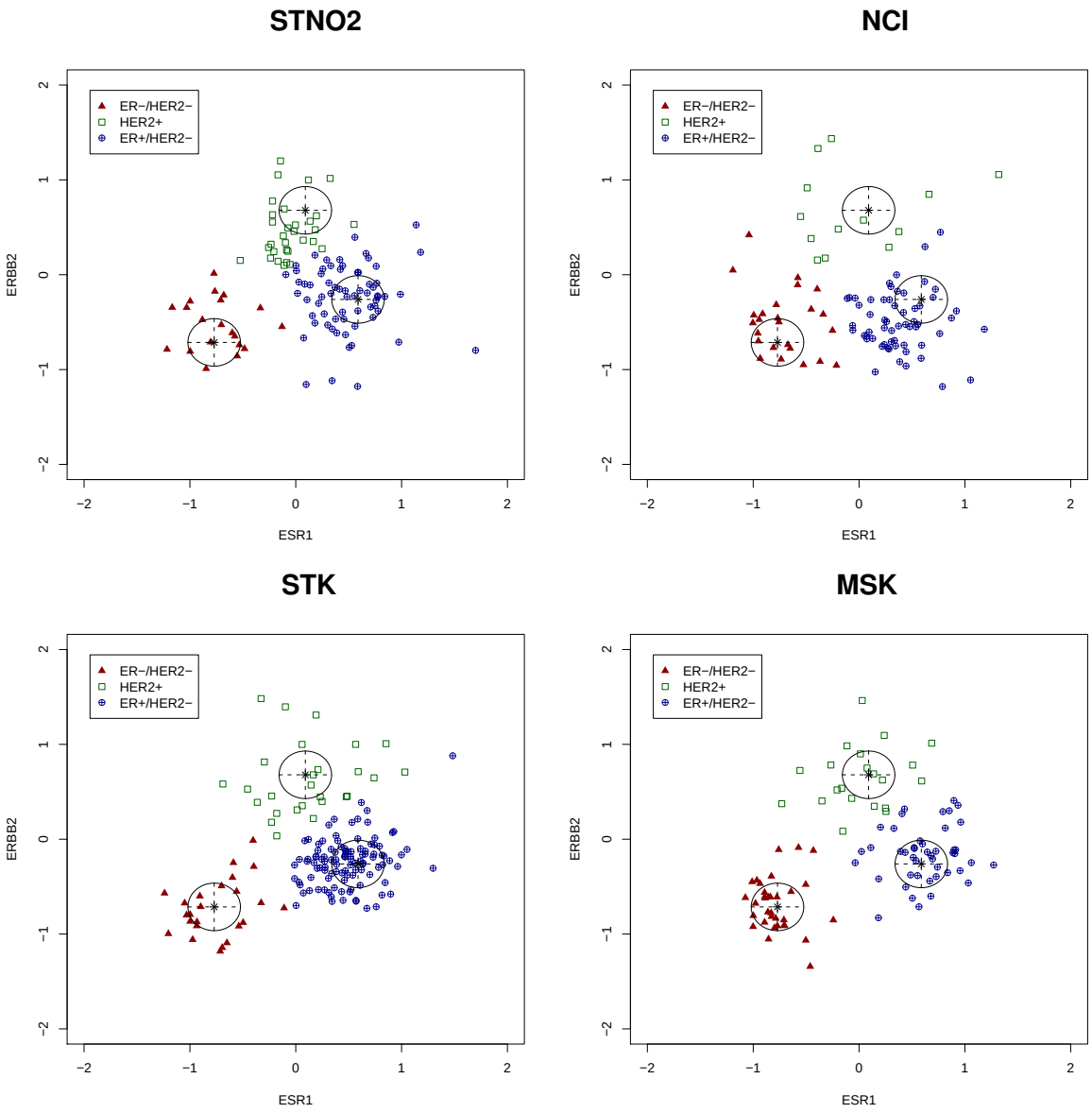


Figure B.2: Classification of the tumors using the subtype clustering model in the STNO2, NCI, STK and MSK datasets.

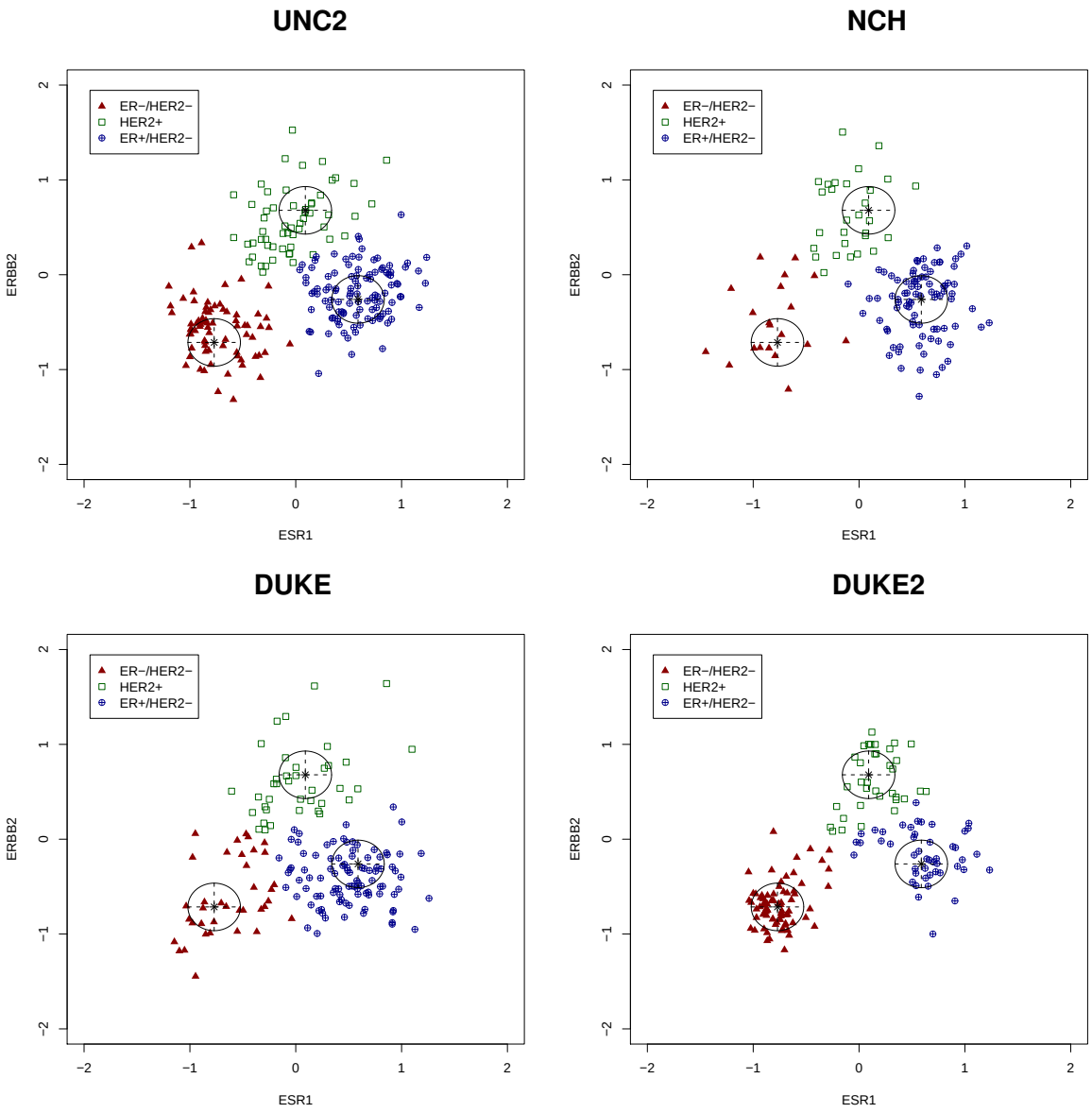


Figure B.3: Classification of the tumors using the subtype clustering model in the UNC2, NCH, DUKE and DUKE2 datasets.

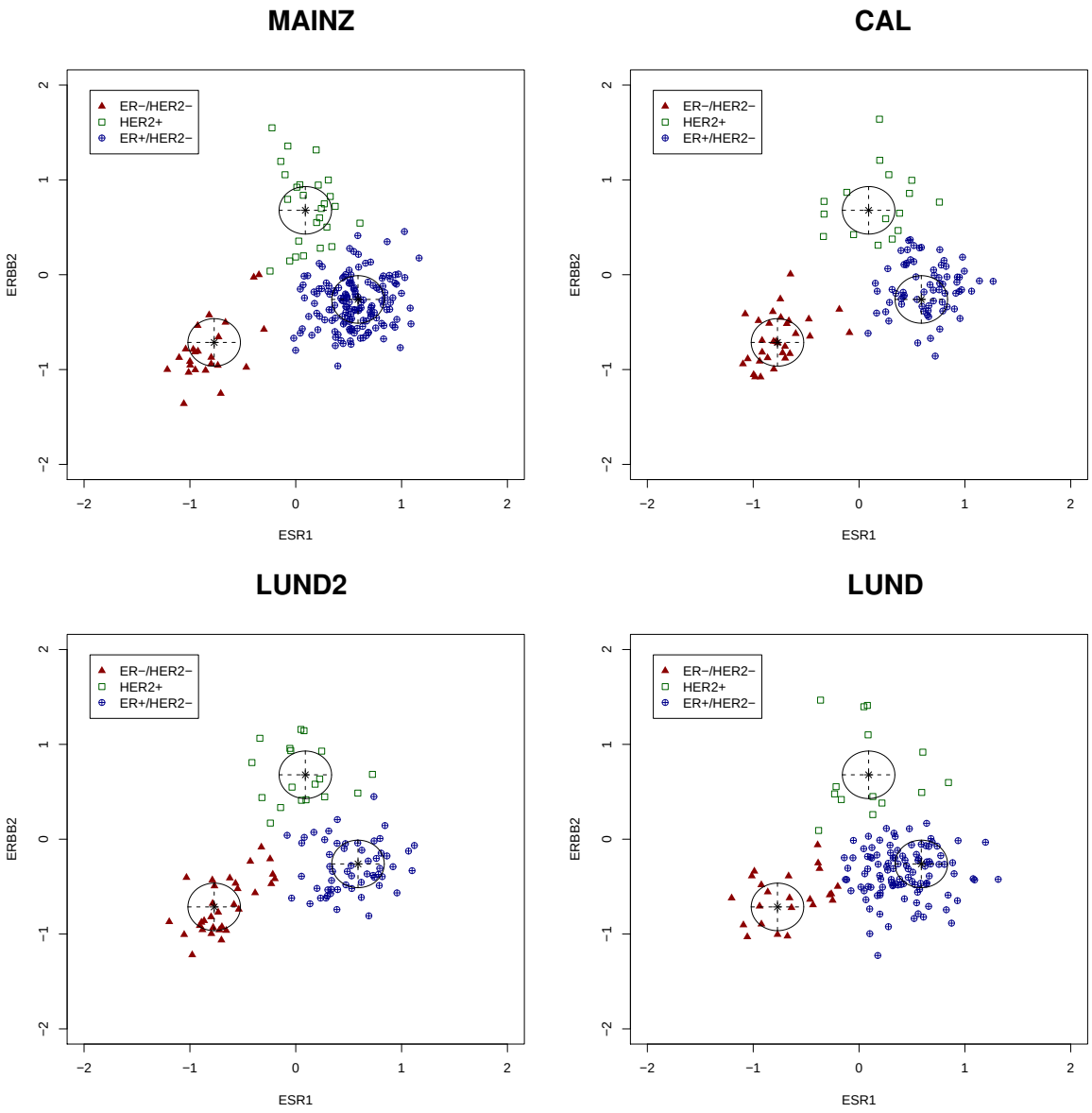


Figure B.4: Classification of the tumors using the subtype clustering model in the MAINZ, CAL, LUND2 and LUND datasets.

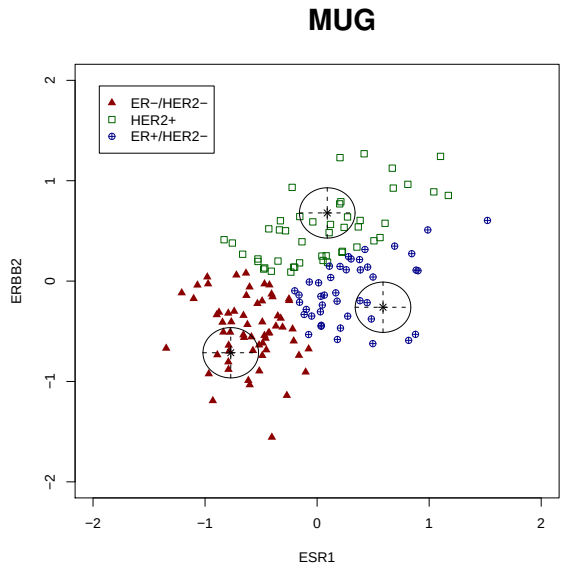


Figure B.5: Classification of the tumors using the subtype clustering model in the MUG dataset.

B.1.1 Perou's Method

Dataset	*ps*			
	2 clusters	3 clusters	4 clusters	5 clusters
NKI	0.89	0.42	0.25	0.20
TBG	0.92	0.39	0.24	0.22
UPP	0.71	0.51	0.33	0.22
UNT	0.78	0.55	0.36	0.00
STNO2	0.86	0.44	0.28	0.25
NCI	0.93	0.44	0.33	0.13
STK	0.61	0.38	0.28	0.25
MSK	0.77	0.66	0.20	0.00
UNC2	0.81	0.57	0.31	0.00
NCH	0.66	0.49	0.36	0.29
DUKE	0.57	0.42	0.37	0.42
DUKE2	0.92	0.63	0.47	0.00
MAINZ	0.68	0.39	0.24	0.18
CAL	0.84	0.41	0.31	0.00
LUND2	0.92	0.51	0.17	0.17
LUND	0.55	0.36	0.24	0.20
MUG	0.50	0.33	0.27	0.23
mean	0.76	0.47	0.30	0.16
sd	0.14	0.10	0.07	0.12

Table B.1: Prediction strength *ps* for Perou's method with respect to the number of clusters (two to five) in the clustering model.

B.2 Local Prognostic Gene Signatures

B.2.1 Gene Modules and Breast Cancer Molecular Subtypes

B.2.1.1 Concordance Indices for Clinical Variables and Gene Modules

Subtype	Variable	Concordance index	95%CI	P-value	n
ALL	age	0.39	[0.32,0.47]	3.8E-03	724
	size	0.64	[0.57,0.71]	5.4E-05	724
	er	0.33	[0.26,0.41]	8.8E-06	718
	grade	0.72	[0.67,0.77]	2.2E-16	708
	ESR1	0.43	[0.39,0.48]	3.5E-03	724
	ERBB2	0.54	[0.49,0.59]	9.5E-02	724
	AURKA	0.67	[0.63,0.71]	6.7E-19	724
	PLAU	0.47	[0.42,0.51]	1.3E-01	724
	VEGF	0.59	[0.54,0.63]	6.7E-05	724
	STAT1	0.49	[0.45,0.53]	6.0E-01	724
	CASP3	0.53	[0.49,0.58]	1.4E-01	724
ER+/HER2-	age	0.39	[0.3,0.49]	2.7E-02	505
	size	0.67	[0.58,0.76]	1.6E-04	505
	er	0.46	[0.22,0.7]	7.2E-01	502
	grade	0.75	[0.68,0.81]	6.1E-13	492
	ESR1	0.51	[0.45,0.57]	6.7E-01	505
	ERBB2	0.58	[0.52,0.64]	8.2E-03	505
	AURKA	0.7	[0.65,0.75]	4.6E-16	505
	PLAU	0.45	[0.39,0.5]	6.3E-02	505
	VEGF	0.57	[0.51,0.63]	1.5E-02	505
	STAT1	0.5	[0.45,0.55]	9.9E-01	505
	CASP3	0.5	[0.44,0.56]	9.3E-01	505
ER-/HER2-	age	0.36	[0.19,0.53]	9.6E-02	115
	size	0.55	[0.39,0.71]	5.6E-01	115
	er	0.24	[0,0.49]	4.1E-02	113
	grade	0.53	[0.35,0.7]	7.7E-01	113
	ESR1	0.49	[0.41,0.58]	8.8E-01	115
	ERBB2	0.54	[0.46,0.63]	3.3E-01	115
	AURKA	0.47	[0.38,0.57]	5.5E-01	115
	PLAU	0.5	[0.41,0.59]	9.9E-01	115
	VEGF	0.54	[0.45,0.62]	4.2E-01	115
	STAT1	0.4	[0.32,0.48]	1.1E-02	115
	CASP3	0.58	[0.5,0.66]	5.6E-02	115
HER2+	age	0.55	[0.38,0.72]	5.6E-01	104
	size	0.56	[0.4,0.73]	4.6E-01	104
	er	0.51	[0.34,0.68]	9.1E-01	103
	grade	0.54	[0.39,0.69]	5.7E-01	103
	ESR1	0.53	[0.44,0.63]	5.0E-01	104
	ERBB2	0.53	[0.44,0.61]	5.2E-01	104
	AURKA	0.57	[0.47,0.66]	1.7E-01	104
	PLAU	0.55	[0.45,0.65]	3.1E-01	104
	VEGF	0.6	[0.51,0.69]	2.3E-02	104
	STAT1	0.36	[0.27,0.45]	3.1E-03	104
	CASP3	0.46	[0.38,0.55]	4.4E-01	104

Table B.2: Concordance indices of the clinical variables and the gene module scores with respect to the breast cancer molecular subtypes.

B.2.1.2 Concordance Indices for Gene Signatures

Subtype	Variable	Concordance index	95%CI	P-value	n
ALL	GENE70	0.68	[0.65,0.72]	3.8E-22	724
	GENE76	0.62	[0.58,0.66]	7.0E-09	724
	P53	0.62	[0.58,0.66]	2.4E-08	724
	WOUND	0.65	[0.61,0.69]	6.4E-14	724
	GGI	0.67	[0.63,0.7]	7.3E-18	724
	ONCOTYPE	0.66	[0.62,0.7]	2.5E-15	724
	IGS	0.62	[0.58,0.66]	2.9E-09	724
ER+/HER2-	GENE70	0.71	[0.65,0.76]	6.7E-15	505
	GENE76	0.63	[0.58,0.69]	3.5E-06	505
	P53	0.59	[0.54,0.65]	1.6E-03	505
	WOUND	0.66	[0.61,0.71]	3.5E-10	505
	GGI	0.7	[0.65,0.75]	1.5E-14	505
	ONCOTYPE	0.68	[0.62,0.73]	1.1E-09	505
	IGS	0.63	[0.57,0.69]	1.8E-05	505
ER-/HER2-	GENE70	0.53	[0.44,0.62]	4.9E-01	115
	GENE76	0.51	[0.42,0.59]	8.3E-01	115
	P53	0.42	[0.33,0.51]	7.6E-02	115
	WOUND	0.51	[0.42,0.6]	8.1E-01	115
	GGI	0.5	[0.41,0.6]	9.3E-01	115
	ONCOTYPE	0.49	[0.39,0.59]	8.3E-01	115
	IGS	0.44	[0.35,0.53]	1.9E-01	115
HER2+	GENE70	0.56	[0.48,0.65]	1.5E-01	104
	GENE76	0.54	[0.45,0.64]	3.8E-01	104
	P53	0.52	[0.43,0.62]	6.3E-01	104
	WOUND	0.58	[0.49,0.67]	8.1E-02	104
	GGI	0.52	[0.43,0.61]	6.3E-01	104
	ONCOTYPE	0.55	[0.46,0.64]	2.9E-01	104
	IGS	0.53	[0.44,0.61]	5.6E-01	104

Table B.3: Concordance indices of the gene signatures with respect to the breast cancer molecular subtypes. GENE70: [van't Veer et al., 2002]; GENE76: [Wang et al., 2005]; P53: [Miller et al., 2005]; WOUND: [Chang et al., 2004]; GGI: [Sotiriou et al., 2006]; ONCOTYPE: [Paik et al., 2004]; IGS: [Liu et al., 2007].

B.2.2 Gene Expression Prognostic Index Using Subtypes (GENIUS)

B.2.2.1 Performance Assessment and Comparison for Gene Prognostic signatures

Subtype	Variable	Concordance index	95%CI	P-value	n
ALL	GENIUS	0.7	[0.67,0.74]	1.0E-27	724
	AURKA	0.67	[0.63,0.71]	4.5E-19	724
	GGI	0.67	[0.63,0.71]	9.1E-19	724
	STAT1	0.51	[0.47,0.55]	3.0E-01	724
	PLAU	0.47	[0.42,0.51]	7.8E-02	724
	IRMODULE	0.58	[0.53,0.63]	4.7E-04	553
	SDPP	0.66	[0.62,0.7]	2.6E-16	724
ER+/HER2-	GENIUS	0.7	[0.65,0.75]	6.9E-16	503
	AURKA	0.7	[0.65,0.75]	1.8E-15	503
	GGI	0.7	[0.64,0.75]	4.3E-14	503
	STAT1	0.51	[0.46,0.57]	3.5E-01	503
	PLAU	0.44	[0.38,0.5]	2.0E-02	503
	IRMODULE	0.6	[0.54,0.67]	1.4E-03	388
	SDPP	0.67	[0.62,0.72]	1.4E-10	503
ER-/HER2-	GENIUS	0.65	[0.57,0.73]	7.1E-05	116
	AURKA	0.47	[0.38,0.57]	3.0E-01	116
	GGI	0.51	[0.41,0.6]	4.3E-01	116
	STAT1	0.6	[0.52,0.68]	5.1E-03	116
	PLAU	0.49	[0.4,0.58]	4.3E-01	116
	IRMODULE	0.63	[0.54,0.71]	2.3E-03	87
	SDPP	0.55	[0.46,0.64]	1.4E-01	116
HER2+	GENIUS	0.65	[0.55,0.74]	9.3E-04	105
	AURKA	0.56	[0.46,0.65]	1.2E-01	105
	GGI	0.52	[0.43,0.61]	2.9E-01	105
	STAT1	0.61	[0.53,0.7]	4.9E-03	105
	PLAU	0.58	[0.49,0.68]	4.9E-02	105
	IRMODULE	0.68	[0.59,0.77]	5.3E-05	78
	SDPP	0.63	[0.55,0.72]	1.4E-03	105

Table B.4: Concordance indices of GENIUS and the prognostic gene signatures with respect to the breast cancer molecular subtypes. AURKA: [Desmedt et al., 2008]; GGI: [Sotiriou et al., 2006]; STAT1: [Desmedt et al., 2008]; PLAU: [Desmedt et al., 2008]; IRMODULE: [Teschendorff et al., 2007]; SDPP: [Finak et al., 2008].

B.2.2.2 Performance Assessment and Comparison for Gene Prognostic signatures

Subtype	Variable	Concordance index	95%CI	P-value	n
ALL	GENIUS	0.7	[0.67,0.74]	1.0E-27	724
	AOL	0.63	[0.59,0.67]	2.5E-11	724
	NPI	0.67	[0.63,0.7]	3.2E-18	708
ER+/HER2-	GENIUS	0.7	[0.65,0.75]	6.9E-16	503
	AOL	0.65	[0.59,0.7]	3.7E-08	503
	NPI	0.69	[0.64,0.74]	4.9E-14	490
ER-/HER2-	GENIUS	0.65	[0.57,0.73]	7.1E-05	116
	AOL	0.54	[0.44,0.63]	2.2E-01	116
	NPI	0.52	[0.43,0.62]	3.1E-01	114
HER2+	GENIUS	0.65	[0.55,0.74]	9.3E-04	105
	AOL	0.56	[0.48,0.64]	6.3E-02	105
	NPI	0.56	[0.48,0.65]	7.7E-02	104

Table B.5: Concordance indices of GENIUS and the prognostic clinical models with respect to the breast cancer molecular subtypes. AOL: [Ravdin et al., 2001]; NPI: [Todd et al., 1987].

Appendix C

Contributive Prognostic Gene Signatures

C.1 GGI

The GGI signature is available from the Supplemental Table 1 in [Sotiriou et al., 2006]. The table is also available from `http://www.ulb.ac.be/di/map/bhaibeka/gene_signatures/sotiriou2006_ggi_signature_128.csv`.

C.2 TAMR13

The TAMR13 signature is available from the Additional File 2 in [Loi et al., 2008]. The table is also available from `http://www.ulb.ac.be/di/map/bhaibeka/gene_signatures/tamr13_genes.csv`.

C.3 Gene Modules

The gene modules are available from the Supplemental Table S1 in [Desmedt et al., 2008]. The table is also available from `http://www.ulb.ac.be/di/map/bhaibeka/gene_signatures/desmedt2008_modules.csv`.

C.4 GENIUS

The GENIUS signature is available from the Supplementary Table 3 in [Haibe-Kains et al., 2009]. The table is also available from `http://www.ulb.ac.be/di/map/bhaibeka/gene_signatures/genius_subtype_model.csv`.